Lessons Learned from Two Cohorts of Personal Informatics Self-Experiments

NEDIYANA DASKALOVA, Brown University, USA
KARTHIK DESINGH, Brown University, USA
ALEXANDRA PAPOUTSAKI, Brown University, USA
DIANE SCHULZE, Brown University, USA
HAN SHA, Brown University, USA
JEFF HUANG, Brown University, USA

Self-experiments allow people to investigate their own individual outcomes from behavior change, often with the aid of personal tracking devices. The challenge is to design scientifically valid self-experiments that can reach conclusive results. In this paper, we aim to understand how novices run self-experiments when they are provided with a structured lesson in experimental design. We conducted a study on self-experimentation with two cohorts of students, where a total of 34 students performed a self-experiment of their choice. In the first cohort, students were given only two restrictions: a specific number of variables to track and a set duration for the study. The findings from this cohort helped us generate concrete guidelines for running a self-experiment, and use them as the format for the next cohort. A second cohort of students used these guidelines to conduct their own self-experiments in a more structured manner. Based on the findings from both cohorts, we propose a set of guidelines for running successful self-experiments that address the pitfalls encountered by students in the study, such as inadequate study design and analysis methods. We also discuss broader implications for future self-experimenters and designers of tools for self-experimentation.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; *Ubiquitous and mobile computing systems and tools*; *Empirical studies in ubiquitous and mobile computing*;

Additional Key Words and Phrases: personal informatics; self-tracking; quantified self

ACM Reference Format:
Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. 2017. Lessons Learned from Two Cohorts of Personal Informatics Self-Experiments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 46 (September 2017), 22 pages.
DOI: http://doi.org/10.1145/3130911

1 INTRODUCTION

"To find out what happens when you change something, it is necessary to change it." —George Box

The current paradigm in research on behavior change as performed in fields such as public health, social sciences, and research initiatives like mHealth [14, 33], is to find generalizable effects that can be disseminated to the public. However, by definition there only has to be a small effect on a subset of the population for those

Author's addresses: N. Daskalova, K. Desingh, A. Papoutsaki, D. Schulze, H. Sha and J. Huang, Computer Science Department, Brown University; contact author: N. Daskalova, (Current address) 115 Waterman Street, Brown University, Providence, Rhode Island, 02912.

2474-9567/2017/9-ART46 $15.00
DOI: http://doi.org/10.1145/3130911

studies to claim a positive result. For example, a general sleep hygiene recommendation to "go to bed earlier" may improve the average productivity across a large study population, but may be detrimental to those with eveningness chronotype [21]. In contrast, when users empower themselves through quantifying aspects of their lives and running their own experiments, they are in essence doing single-subject science. Therefore, a personalized approach to experimentation is more relevant. The goal of personal informatics is not to discover knowledge about a broad population, but to help people learn about what affects them, specifically for the parts of their lives that matter to them the most.

Personal informatics incorporates the collection, analysis, and reflection on various facets of personal data and experiences, primarily with the aid of technology. Recent research has shed light on the different directions of self-tracking: it has identified how people perform self-tracking, what reasons motivate them to do so, and what data they track most commonly [7, 31]. The findings from these investigations have shown that people generally perform descriptive analyses and that tracking one's own behavior is a beneficial process. People perform self-tracking for various reasons, including to be mindful of their behavior or to improve their lives by solving a current problem they are experiencing. Although self-tracking has other benefits and can reveal interesting insights or correlations, we focus on determining a causal relationship, which requires a different approach: self-experimentation. Our findings are aimed at empowering people to run effective self-experiments.

Beyond passive monitoring, the next step towards better understanding one's self is to perform self-experiments: to create and test hypotheses on the effect of small behavior changes [15]. In self-experiments, people can be motivated to find a behavior change that would alleviate a given problem that they are encountering. However, many individuals who perform self-tracking do not have the capability to conduct analyses or run rigorous experiments, and may create "under-specified goals that were [are] not actionable" [30]. In this paper we seek to answer: (1) what lessons for future self-experiments can we extract from observing novices run self-experiments, and (2) how do guidelines on self-experiment design affect the way people run self-experiments and analyze their data?

We present the findings on self-experimentation from of two cohorts of participants (a total of 34 students in a Human-Computer Interaction seminar) performed an experiment of their choice on themselves as part of a class assignment. The students designed hypotheses, tracked the appropriate variables, and submitted reports comprising their procedures, a day-by-day journal, visualizations, and analyses. The first cohort was given minimal guidance, and the lessons we learned from how they conducted their self-experiments informed a structured set of self-experiment guidelines. In order to evaluate their effectiveness, we then asked the second cohort of participants to follow these guidelines. Finally, we further iterated on the guidelines based on what we learned from the second cohort's self-experiments.

There are two main contributions of this work: (1) a series of lessons about self-experimentation from an exploratory study with two cohorts performed in a classroom setting, and (2) a proposed set of guidelines which aims to help non-scientists run N-of-1 style self-experiments and discover positive effects of behavior change. This is the first systematic analysis of multiple self-experiments conducted in a structured and guided environment where participants are given the freedom to choose their experiment. We combine the findings from both cohorts to present these guidelines, which can help both future self-experimenters and designers of self-experimentation tools, as one way of conducting an iterative self-experiment.

2 RELATED WORK

2.1 Self-Experimentation

Self-experimentation is a type of scientific experiment in which the experimenter herself is the only subject involved [41]. Sanctorius of Padua is one of the earliest documented examples of this type of research. In the early 17th century, Sanctorius weighed himself daily over a period of thirty years along with his food, liquid

intake, and body excretions. This self-experiment led to the discovery of metabolism [12]. In our study, some students also logged their weight, but with electronic devices, which simplified their data analysis.

Since then, self-experimentation has been applied in a variety of research fields such as medicine and psychology. Self-experiments take various forms: some are momentary while others are longer-term [34]. One notable example of self-experiments performed as a classroom assignment is that of Allen Neuringer in an introductory psychology class [32]. Neuringer asked his students to perform a self-experiment for 2 weeks to illustrate the possibilities of self-experiments outside the laboratory. Some of his students designed their experiments in phases and mainly looked at the difference of means between the conditions. We build on that model by providing more structure to the assignment and increasing the duration of the experiment.

2.2 Personal Informatics

People have been tracking their own behavior, health, and feelings for a long time. Diaries are an example of such record-keeping as they provide the means to look back and reflect on one's experiences, or simply because "we forget all too soon the things we thought we would never forget" [9]. Recently, self-tracking devices have become ubiquitous, allowing passive tracking of a wide range of variables. People can now record not only aspects of their health such as calories and amount of time slept, but also how they spend their time and money. The most common tools for self-tracking are smartphones and other portable and wearable devices, like the popular FitBit [1]. However, the amount of data people collect about themselves is so overwhelming that certain innovations focus on synthesizing the information from multiple platforms and presenting it in a simpler, more understandable form [5].

Hekler et al. point out that many current self-tracking technologies do not provide the tools to self-experiment, as knowledge on its own is not enough for behavior changes [19]. The key to forming new habits is to provide context that links new behaviors to existing ones. Fogg emphasizes in his "Three Tiny Habits System" the importance of fostering new habits by starting with small, incremental steps based on established behavioral routines [15]. This implies that knowing what change to make is not always enough to actually change behavior. However, they can more easily create a new habit as they are already seeing positive changes from their self-experiment.

The "Quantified Self" community comprises individuals who use and design tools for personal informatics [28]. Quantified-Self participants hold Meetups around the world, during which they present what they tracked, how they tracked it, and what they learned from it. Choe et al. studied this community by analyzing videos from the Meetups and extracting valuable lessons from the self-tracking practices of this extreme user group [7]. They found that Quantified-Self enthusiasts compromised the validity of their results due to three common pitfalls: 1) tracking too many things, 2) not tracking triggers and context, and 3) lacking scientific rigor, such as not including control conditions. In our research, we look at how guidance can support more scientifically rigorous N-of-1 style experiments.

2.3 Self-Experimentation for Personal Informatics

The value of personal informatics comes from the process of discovery and reflection on one's data. Anyone can start self-tracking, but only people who know what to study and how to interpret the results will gain useful insights [36]. Li et al. derived a stage-based model of personal informatics composed of five stages (preparation, collection, integration, reflection, and action) and identified barriers that current systems pose in each stage [31]. They argue that personal informatics tools should allow users to iterate on their experimental stages to find the optimal procedure; this supports the iterative model of self-experiment design. Epstein et al. [13] build on Li et al.'s model by expanding in preparing and selecting tools for behavior change and in maintenance. They emphasize that some trackers either wanted to or had to change tools during their experiments [13]. That supports our

finding that there is a need for a phase during which experimenters can explore various tools and actually try them in the context of their experiment.

Some forms of personal informatics do not require self-experimentation. For instance, people track simple things such as daily steps or number of push-ups to motivate themselves toward specific goals, or to archive various aspects of their lives as a new way of journaling and reflection [7]. However, to determine if there are causal relationships between variables in their lives, people must perform self-experimentation to get scientifically valid results. Self-experiments can be further motivated by a user trying to solve a problem by finding the right behavior change protocol to address the issue [30]. Lee et al.'s work investigate how to help people (mostly students) develop a behavior change protocol using habits developed using triggers and SMART (Specific, Measurable, Actionable, Realistic, and Timely) goals. We also look at the next step of having the students develop N-of-1 trials, which begin after 5 weeks, similar to Lee et al.'s guidance of "incorporat[ing] n-of-1 trials after the week 5 or 7 mark." Analysis of these N-of-1 style experiments is a challenge, and Section 6.6 proposes some computer-aided Bayesian solutions.

Roberts has played a pioneering role in introducing self-experimentation to the self-trackers that are new to the Quantified Self community [34, 35]. He ran numerous experiments over a period of twelve years, identifying several novel causal relationships which he later found to be related to conventional research findings. He also popularized a method for weight reduction based on his experiments, which was anecdotally reported to be effective. He argues that self-experimentation has several benefits over conventional research, including strong self-motivation, no limit to experiment duration, and easier idea generation and validation.

Karkar et al. present an initial framework of how to run self-experimentation, specifically with a focus on problems with irritable bowel syndrome [25]. While the framework works well for the proposed cases with no carryover effect, it would need to be customized further to adapt for other domains. Furthermore, another limitation of the main premise of the framework is that the self-experimenter would re-run the study if they are not satisfied with the results after the end of the study. We build on this work by accounting for these issues in our proposed guidelines.

2.4 Single-Case Experimental Design

In order to find an appropriate single-case intervention design, which reduces the confounding effects of time-based events, we turn to single-case design literature describing N-of-1 trials: specifically to Kratochwill et al.'s standards for single-case designs [27]. These standards, created by a group of quantitative and single-case design methods experts, are based on the AB phase design. In the AB experimental design, the participants change their behavior at a predefined time in order to analyze the effects of the independent variable on a dependent variable [40]. This change in behavior signifies a new phase of the experiment. The A phases are one pattern of behavior, usually the baselines, and the B phases are another pattern of behavior, usually the intervention. Figure 1 shows the different design methods and their phases in a time series experiment [3]. For example, the ABA experimental method divides the experimental period into three phases: the experimenter starts with behavior pattern A, followed by behavior pattern B, and then A again.

According to Kratochwill et al., the ABAB phase design variation is the most appropriate one, as the basic AB design is susceptible to confounding variables, thus affecting the quality of the conclusions [27]. The ABAB design, and even more sophisticated designs such as the ABABAB, allows for the suggested minimum of three attempts of change to demonstrate the intervention effect. The ABAB phase design, which combines the least bias with enough time for users to acclimate to the recommendations, was used in a recent research project, that studied the success of a new sleep tracking system, SleepCoacher [8], that provided participants with personalized sleep recommendations.

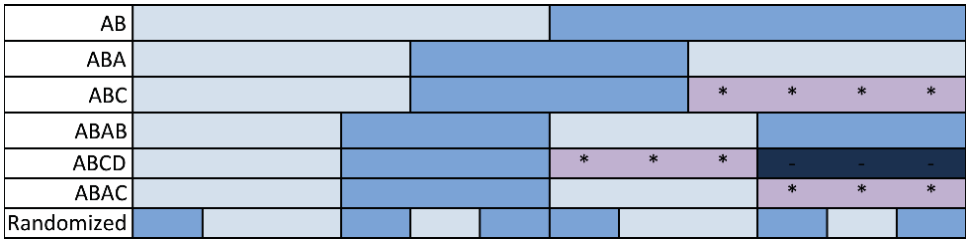

Fig. 1. Experimental design methodologies can have different time series patterns. For the randomized design, each phase is just one measurement. For AB designs, each phase has a randomized length with multiple measurements. There can also be more than just two types of phases: in ABC, for example, there are two levels of treatment in addition to the control phase.

Researchers such as Choe et al. [7] have compiled tips for novice self-trackers based on advice given by experienced ones. In this paper, we describe guidelines for self-experimentation that we developed based on a systematic analysis of one cohort of self-experimenters. We then evaluate these guidelines with a second cohort of self-experimenters. Finally, we iterate on our initial guidelines and outline their improved version to help designers of self-experimentation tools and future self-experimenters who are looking for guidance.

3 SELF-EXPERIMENT STUDY

3.1 Study Method

We performed an exploratory study with two cohorts of participants: Cohort 1 and Cohort 2. The purpose of the study with Cohort 1 was to learn how novices perform self-experiments when given minimal guidance. Based on those findings, we outlined a series of guidelines that can be used to conduct successful self-experiments. We then evaluated the success of those guidelines on novice self-experimenters with Cohort 2. The study in both cohorts was distributed as an assignment in two offerings of a Human-Computer Interaction seminar at a university.

We used directed content analysis to analyze the students' reports [22]. Two coders read the reports individually and coded them according to the original codes. When there was a disagreement, the coders re-read the report and reached a compromise together. The main original codes matched the main sections that each student was supposed to include in their report, such as: descriptions of their independent and dependent variables, hypotheses, tracking methods, reasons for choosing all of those, as well as analysis methods, results and conclusions, challenges they faced, and what they would have done differently if they were to do it again. The "challenges" code was broken down further to match the pitfalls outlined by Choe et al. [7], and new codes were created for the ones that did not match the original pitfalls.

Based on the findings from Choe et al.'s study [7], self-experimenters need some background in analysis and visualizations in order to effectively design their own experiments and learn from the results. Students in our cohorts developed the essential background by reading about analysis methods and visualizations and discussing them in class. They were also introduced to topics in experimental methods and behavior analysis. However, building on Choe et al.'s advice, we learned that it is not enough to have a basic understanding of analysis methods; rather, self-experimenters need to be cognizant of specific methods for analyzing results from single-subject experiments.

3.2 Students' Self-Experimentation Methods

3.2.1 Self-Experiments Method: Cohort 1. In the first part of the study, we distributed an assignment in a Human-Computer Interaction seminar where a cohort of 20 undergraduate and graduate Computer Science students (8 female) ran a month-long self-experiment. While there were 21 students in the class, one did not

Table 1. Demographics and experience of students in the two cohorts.

	Cohort 1	Cohort 2
Total Cohort Size (N)	20	14
Male	12	4
Female	8	10
Undergraduate	8	10
Graduate	12	4
Computer Science Concentrators	19	12
Statistics Experience	13	8
Self-Tracking Experience	5	7

Stage	Goal	Length
Stage 1	Exploration	1 week
Stage 2	Preliminary Hypothesis Testing	2 weeks
Stage 3	Real Experiment	6 weeks

Fig. 2. Stages of the study design in Cohort 2: students start with an Exploration period, followed by a Preliminary Hypothesis Testing, and finally they run the Real Experiment for 6 weeks.

consent to disclose their collected data and did not complete the surveys offered at the end of the semester, so we omitted them from the background examination of the experimenters. A summary of the demographics and experience of the students is presented in Table 1.

In total, the assignment lasted 5 weeks, with a combined 1 week for planning and analysis and 4 weeks for tracking. The students were instructed to design and conduct a self-experiment by forming two hypotheses based on at least one independent and two dependent variables. No two experiments could be the same, but this did not seem to constrain the students' choices, as can be seen from the broad range of experiments shown in Table 2.

The students in Cohort 1 were not given any direct guidance on exactly what type of analysis to perform, and some only used visual analysis. We further discuss their results in the Findings section.

3.2.2 Self-Experiments Method: Cohort 2. In the second part of the study, we distributed an assignment in a later offering of a Human-Computer Interaction seminar, where a cohort of 15 undergraduate and graduate students ran a semester-long self-experiment. They all consented to the use of their data for this study. There were 15 students (10 female) in the class, but one did not submit a final report, so we omitted them from the background examination of the experimenters. Two of the students in Cohort 1 also took the seminar in its second offering and they are also part of Cohort 2.

The assignment was run in three stages, as shown in Figure 2. In Stage 1, Exploration, students simply tracked anything they wanted. The goal was to explore many tracking tools and variables so they could familiarize themselves with what was possible. In Stage 2, Preliminary Hypothesis Testing, students narrowed their interest by creating specific hypotheses they wanted to test and by tracking the related variables for 2 weeks: one week where they just tracked, and a second week during which they introduced an intervention. At the end of the two weeks, they performed statistical analysis on the data. The goal of this stage was to introduce students to running a short version of the experiment and uncover any potential future problems with data collection or analysis. Therefore, before their official experiments began, they had the chance to think about their setup and

iterate on the design, hypotheses, and variables if necessary. In Stage 3, Real Experiment, students performed a randomized ABAB self-experiment for a duration of 6 weeks and analyzed their data.

In total, the assignment lasted 9 weeks. Similar to Cohort 1, Cohort 2 students were instructed to design and conduct a self-experiment by forming two hypotheses based on at least one independent and two dependent variables. No two experiments could be the same but again this did not constrain their choices.

Data from self-experiments is autocorrelated and probably not normally distributed, which means that not all regular statistical analysis methods are appropriate. Participants in Cohort 2 were asked to perform specific analysis on their collected data: a t-test and an effect size calculation with a confidence interval for the standardized mean difference. According to single-case design literature, standardized mean differences and effect sizes are appropriate for self-experimenters as they are simple to perform [39]. We further discuss reasons for choosing these approaches in the "Participants' Expertise with Statistics and Personal Informatics" subsection of the Discussion.

Using a t-test, Cohort 2 students considered the hypothesis conclusive if the relationship was statistically significant ($p < 0.05$) and inconclusive otherwise. In an attempt to balance the understanding of our participants with what they can feasibly learn from these self-experiments, we labeled statistically significant tests as "conclusive." We recognize, however, that this is technically not the case and there is a lot more subtlety involved in these statistics, which we discuss later in the paper. A conclusive result suggests that the null hypothesis of the experiment should be rejected, as the relationship between the variables is not likely to be random. On the other hand, an inconclusive result suggests that the observed effects may be a consequence of randomness rather than causation. Students also computed the standardized mean differences and Hedge's g for effect size, along with its confidence interval. A 95% confidence interval for Hedge's g suggests with 95% confidence the result will be in this range if the experiment is repeated.

3.2.3 Tracking and Measuring Self-Experiment Variables. In both cohorts, students were encouraged to use computing devices to monitor the variables since the study would last a few months in total. Most students used smartphones and wearable devices, either their own or from a loan pool of wearable devices that we provided. Others logged their observations and measurements on spreadsheets. At the end of the experiment, each student submitted a report describing their hypotheses, variables tracked, statistics used, a day-by-day journal (both textual and numeric), and visualizations and analyses performed to test the hypotheses. The assignment also asked students to include a discussion of lessons learned and whether the results matched their expectations.

Students were instructed to track any combination of independent and dependent variables, as long as there was a testable hypothesis and the data could be analyzed. We observed commonalities across all students in the various aspects of the process, including variables, confounding factors, and statistical results.

3.3 Participants' Expertise with Statistics and Personal Informatics

The students in both cohorts are part of a specific population that is capable of quickly learning scientific methods and generating visualizations and analyses for their experiments. However, the students also had varied experience in experimental design. Upon completion of the class, all students were asked to complete a survey asking them about their level of expertise with statistics and personal informatics. In Cohort 1, most students had statistical background before taking this class (13 of 20 students), but the majority of them (15 of 20) lacked previous experience with self-tracking. Based on an informal poll, none of them had previous experience with self-experiments, which is what classified them as "novice self-experimenters". In order to ensure that all students had a foundation in personal informatics, information on statistical methods and personal informatics was disseminated through paper readings and discussions on experimental methods and behavioral analysis prior to the study.

Table 2. Self-Experiments of Students in Cohort 1 (* - visual analysis, Result - whether the result matched the expectation of the self-experimenter).

ID	Independent Variable(s)	Dependent Variable(s)	Hypotheses	Design	Result
P1	exercised for 30 minutes	sleep quality	more exercise, better sleep quality	AB	YES
	number of steps	sleep quality	more steps, better sleep quality		Inconclusive
	type of task	heart rate	heart rate differs between tasks		YES
P2	number of classes attended	productivity	less attendance, more productivity	ABA	Inconclusive
		time spent online shopping	less attendance, less time spent online shopping		YES
		unnecessary spending	less attendance, less unnecessary spending		YES
P3	weather conditions	exercise frequency	better weather conditions, more frequent exercise	Naturalistic	Inconclusive
		exercise duration	better weather conditions, longer exercise		Inconclusive
P4	ran before or after 6pm	weight	running before 6pm does not increase weight loss	ABA	YES*
		average pace	running before 6pm does not lower average pace		YES*
P5	amount of green tea consumed	times woken up	more tea, less times woken up	ABAC	NO
		time spent sleeping	more tea, more time spent sleeping		Inconclusive
		mood	more tea, better mood		Inconclusive
P6	showered before bed	time to fall asleep	shower before bed, less time to fall asleep	Randomized	Inconclusive
		resting heart rate	shower before bed, lower resting heart rate		Inconclusive
		amount of restful sleep	shower before bed, more restful sleep		Inconclusive
P7	sensitivity/reactivity to food	weight loss	avoiding reactive foods increases rate of weight loss	ABAB	YES
		mood, stress, energy & body feel	avoiding reactive foods improves overall well-being		YES
P8	exercised, took supplements	weight	more exercise & supplements, more weight		YES*
		heart rate	more exercise & supplements, higher heart rate	AB	YES*
		oxygen saturation in blood	more exercise & supplements, higher oxygen saturation		NO*
		stress levels	more exercise & supplements, less stress		NO*
		sleep quality	more exercise & supplements, better sleep quality		NO*
P9	electronics used past 9pm	sleep quality	no electronics after 9pm, better sleep quality	ABA	Inconclusive
		type and intensity of dream	no electronics after 9pm, no effect on dreams		YES*
		dream content and recall	no electronics after 9pm, no effect on dreams		YES*
P10	ran for 30 mins	sleep quality	running affects sleep quality (significantly)	AB	Inconclusive
		heart rate upon waking up	running affects heart rate (significantly)		Inconclusive
		sleep quality	running and sleep are not independent		YES
		steps in bed	running affects steps in bed (significantly)		Inconclusive
P11	minutes being tickled	money spent per week	tickling will increase money spending	ABCD	Inconclusive
		morning mood	tickling will improve mood		Inconclusive
		sleep quality	tickling will improve sleep quality		Inconclusive
		weight	tickling will increase weight		Inconclusive
P12	drank apple cider vinegar	ph level	drinking apple cider vinegar, higher body ph level	AB	YES
		% of time asleep	drinking apple cider vinegar, better sleep quality		Inconclusive
		number of awakenings	drinking apple cider vinegar, better sleep quality		Inconclusive
		time to fall asleep	drinking apple cider vinegar, better sleep quality		Inconclusive
P13	amount of coffee consumed	productivity	more coffee, improved productivity	ABCD	Inconclusive
		sleep	more coffee, less sleep		Inconclusive
P14	ran in the morning	heart rate	morning runs reconcile midday and morning heart rates	ABAB	Inconclusive
		heart rate	morning runs reconcile midday and evening heart rates		Inconclusive
		daily PSS score (stress)	leads to lower total PSS score		Inconclusive
P15	amount of screen time per day	sleep quality	the more screen time per day, poorer sleep	Naturalistic	Inconclusive
	screen-less time before bed	sleep quality	looking at a computer at bed time, poorer sleep		Inconclusive
P16	mean daily temperature	hot beverage drank in the day	the lower the temp, the more hot beverages drank	Naturalistic	YES
		self-report feeling of laziness	the lower the temp, the lazier about working		Inconclusive
P17	amount of smartphone usage	productivity	less phone usage in work hours, more productivity	AB	YES*
		activeness	mobile phone usage affects activeness		NO*
		sleep	mobile phone usage affects sleep		NO*
P18	used time blocking	mood	time blocking will improve mood	AB	Inconclusive
		sleep quality	time blocking will improve sleep quality		Inconclusive
		productivity	time blocking will improve productivity		YES
P19	went swimming for 1.5 hrs	sleep quality	regularly swimming, better sleep quality	ABA	YES
		weight	regularly swimming, reduce weight		YES
		productivity	regularly swimming, improve productivity		NO
P20	consistency of bed/wake time	sleep quality	fixed sleep time window, better sleep quality	ABAC	Inconclusive
		productivity	fixed sleep time window, increase productivity		Inconclusive
		tiredness	fixed sleep time window, reduced tiredness levels		YES

Based on what we learned from the Cohort 1 self-experiments, we created a set of guidelines that novices could use if they were looking for guidance on running such experiments, including how to analyze the results. We presented these guidelines to Cohort 2 participants so that we could evaluate their effectiveness and improve them further. Unlike Cohort 1 students who were introduced to a wide variety of statistical methods, Cohort 2 students focused on using a difference of means test and looked at the size of the effect. We turned to literature about single-case designs and self-experiments for advice on the best way to perform analysis on self-experiment data since this data is autocorrelated and might not be normally distributed. These stipulations are important as they violate the assumptions of most common analysis methods. Despite the abundance of techniques that attempt to address these pitfalls, there exists no single completely agreed-upon method for analyzing data from self-experiments. However, Smith's review of current methods and standards for analysis suggests using standardized mean difference approaches because the effect sizes calculated with these were the least affected by autocorrelation [39]. The advantage of such methods of analysis is that they are relatively simple to perform, which makes them appropriate for novice self-experimenters.

4 SELF-EXPERIMENT OUTCOMES FOR STUDENTS

Since the students were not restricted to a specific set of variables they could track and observe, there was a wide range of hypotheses in both cohorts. Table 2 shows the list of independent and dependent variables chosen by the participants in Cohort 1, along with their hypotheses and experimental outcomes. We refer to the individual student participants as P*n*, where *n* is the participant's ID.

We include Table 2 to show the diverse list of dependent variables that were tracked across all participants in Cohort 1, such as heart rate during different times of the day, productivity, mood, stress, weight, and various sleep variables. There is also great variation in independent variables: amount of coffee consumed, sensitivity to food, number of classes attended, etc. The variables participants tracked in Cohort 2 were just as diverse.

Participants in Cohort 1 used different experimental designs and methods of tracking and measuring variables as per their convenience and available resources. Of the 20 participants, 16 used some form of AB phase design for their experiments, 3 used a naturalistic design (in which they went along with their lives and at the end looked back and calculated correlations), and 1 used only randomization without any kind of phase design. Of the 14 participants in Cohort 2, 12 used some form of AB phase design and 2 used only randomization.

Dependent variables can be divided into two categories based on how they are affected by a particular independent variable. Some dependent variables change gradually, while others change immediately. For example, mood (which is subjectively measured) might be a dependent variable that changes immediately after the independent variable changes, while sleep quality might be changing gradually. The amount of lag and carryover effect in the dependent variable is an important experimental design consideration.

5 STUDY FINDINGS

This section is separated into subsections, each focusing on a single issue that we uncovered in the Cohort 1 study and then addressed with a guideline in the Cohort 2 study. We also report our findings on the effectiveness of the guidelines.

5.1 Iterating on the self-experiment design and selecting variables

A common theme that emerged from the final presentations of the Cohort 1 students was that they wished they had had time to restart their experiments when they realized that they were not collecting data the appropriate way, or if they thought of a better way to measure a variable. For example, one of the main challenges in the Cohort 1 study was that some students did not chose appropriate independent and dependent variables at the beginning of the study. Thus, either their results were affected by confounding factors, or the dependent variable

Table 3. Summary of hypotheses' results from different experimental designs chosen by the Cohort 1 participants.

Design	Conclusive		Inconclusive	Total
	Statistical Analysis	Visual Analysis		
AB	5	8	9	22
ABA	5	4	2	11
ABAB	2	0	3	5
ABAC	2	0	4	6
ABCD	0	0	6	6
Randomized	0	0	3	3
Naturalistic	1	0	5	6

Table 4. Summary of hypotheses' results from different experimental designs chosen by the Cohort 2 participants.

Design	Conclusive	Inconclusive	Total
ABAB	28	29	57
ABABAB	2	0	2
Randomized	1	2	3

was within their control and therefore was biased by their own motivations throughout the study. Furthermore, many students tracked productivity for which there is no standard metric for measuring or tracking it, thus their definitions evolved during the study making the analysis inconsistent. Another challenge in measurement and tracking in the Cohort 1 study was that some variables, such as weight, fluctuate throughout the day and so the time of the measurement affected their findings. As a possible solution, participants could record the variable multiple times per day in order to visualize the variation of data.

These issues informed a fundamental change to the procedure in Cohort 2. In order to address the problem of how to select the best variables and how to best measure and track them, we introduced the idea of an iterative self-experiment to the students in Cohort 2. According to this new guideline, we split the experimental timeline into three stages, as explained earlier in the "Structure of the Self-Experiments: Cohort 2" section: (1) Exploration, (2) Preliminary Hypothesis Testing, and (3) Real Experiment. Thus, students in Cohort 2 had the opportunity to iterate on their variables, devices, and hypotheses before running the full-length experiment.

Students found this iterative process useful: the preliminary stages helped them think of ways to choose variables and optimize their tracking methods. Furthermore, we found that this process decreased the number of inconsistent rating scales and measurements used because it provided time and opportunity for reflection.

5.2 Appropriate self-experiments design

According to Choe et al. [7], the lack of scientific rigor is a common pitfall when conducting self-experiments. In the Cohort 1 study, we avoid this pitfall by introducing participants to a variety of techniques for conducting and analyzing an experiment. During class, students read papers about experimental design methods, causality, validity, interpreting results, and biases. This gave them a basic introduction to help them design their own experiments. However, similarly to what Choe et al. found, students' experiments still lacked scientific rigor and were flawed due to confounding effects.

For Cohort 1, we chose to not give specific guidelines about the design of the experiment beyond the number of variables and hypotheses, which provided a closer look of how novices perform self-experiments. Table 2

illustrates the diversity of AB phase designs used. We refrained from presenting Cohort 1 students with a single design as the best pattern to use because design methodologies are subjective and heavily dependent on the variables chosen. For example, if we measure the effect of "Running 5 miles" on "Mood right after run", then daily randomization might be better than an AB phase design. But if the participant is trying to measure the effect on "Sleep quality", then an AB phase design is preferred as the carryover effect might be large enough to affect more than just one night of sleep. Therefore, the carryover effect can determine whether AB phase design is appropriate.

The students in Cohort 1 chose a variety of experimental design patterns and had similarly varying levels of success. In Table 3, the results are consolidated with respect to these design patterns. Participants predominantly chose the AB and ABA patterns. Design patterns such as ABAC and ABCD are useful because they allow the experimenter to vary the levels of the independent variable. However, the duration of the total experiment should be large enough to sufficiently explore these variations.

In order to find an appropriate single-case intervention design, which reduces the confounding effects of time-based events, we turned to Kratochwill et al.'s standards for single-case designs [27]. Thus, in the Cohort 2 study, one of the guidelines we provided students was to run an ABAB phase design, where the A phases were the baseline days and the B phases were the intervention days. Furthermore, Kratochwill et al. emphasize the need for at least 5 measurements in each phase [27]. If there is only one measurement per day, the experiment needs to be at least 20 days long. However, as we learned in Cohort 1, measurements can be lost due to inappropriate data collection or a variety of other reasons, so we chose to double this minimum length of the self-experiment. Furthermore, the longer experiment also allowed us to introduce randomization of the moment of phase change, which we discuss in the next subsection.

Table 4 summarizes the results of the Cohort 2 self-experiments with respect to design pattern. Students conformed to the suggested designs, and a higher percentage of hypotheses were deemed conclusive in Cohort 2 than in Cohort 1: 75% vs 50%. The reversal design nature of the randomized ABAB design helped decrease the confounding effect of time-based events. It also helped maintain a high level of scientific rigor by following the standards for single-case design interventions.

5.3 Randomization in the self-experiment

The students in Cohort 1 were introduced to randomization as a good practice for an experiment. Randomized single-subject experiments can be helpful for individualized treatments of patients, and systematic replication can lead to insights about a larger population [11]. These experiments use randomized tests to assess the efficacy of a randomly assigned treatment. A major concern among researchers who oppose randomized single-subject experiments is that the treatment might harm the subject if administered randomly [11]. However, since experimenters are also the subjects in self-experiments, they can stop the experiment at any time if they feel any discomfort.

Only P6 in Cohort 1 incorporated randomization in her design methodology. P6's independent variable was whether or not she showered before going to bed at night; she made this decision with a coin toss. P6 knew before the experiment began that she would be traveling across time zones for 10 days of the study, so she wanted to avoid being biased by jet lag in her choice of whether to shower before bed or not. Furthermore, according to her report, she chose the coin flip so that her "fatigue at the end of the day does not affect [her] choice whether to shower at night or not, and [she] let the randomness of the coin decide." This kind of randomized condition is preferred if the carryover effect on the dependent variables is minimal and if the goal is to reduce the possible bias from other environmental factors.

Although randomization was introduced in class as a way to improve the internal validity of experiments in general, no student from the Cohort 1 study who performed some form of AB phase design randomized the start

of each phase. This lack of randomization decreased the validity of their experiments. We believe that students did not realize that randomization could and should be applied to self-experiments specifically, or perhaps they did not know how to apply it to their design. To address this problem in the next study, we provided Cohort 2 students with specific guidelines on how to introduce randomization in their experiments by randomizing the moment of phase change [20].

All students' experiments in the Cohort 2 study involved some randomization. Two of the students performed similar randomization to P6 from Cohort 1: they tossed a coin to decide whether to shower before bed or whether to sniff lavender oil before bed, respectively. These two students' hypotheses both evaluated the effect of the independent variable on the time it took them to fall asleep right after. Therefore, no carryover effect was expected, so they did not need to follow the ABAB phase design. All the other students in the class randomized the moment of phase change by running a script provided by the researchers [20]. The script took as an input the number of total measurements e.g. 8, the number of phases required e.g. 4, the type of design e.g. "ABAB", and the minimum number of measurements in each phase e.g. 5. The output was a string where each character represented the phase of each corresponding measurement, e.g. "AABAAABB."

We find that providing students with a simple script helped them schedule exactly when to change phases. Furthermore, it introduced randomization in their experiments, which helped decrease the effect of confounding variables [20].

5.4 Length of the self-experiment

Another challenge was that Cohort 1 students ran the experiment for only four weeks, which led to an insufficient number of data points and therefore hindered the observation of statistically significant relationships between variables. The last two weeks of the study were also unusual weeks (midterm week and spring break), which may have introduced confounding variables and disruptions to tracking. Many students reported that if they were to do the study again, they would extend the duration of their experiments. According to Kratochwill et al., single-case designs should have at least 4 phases and 5 measurements in each phase [27]. Thus, a suggestion for future self-experimenters is that measuring the dependent variable only once per day (for example, the amount of water intake to evaluate its effect on sleep quality) requires at least 20 days of self-experimentation.

Furthermore, life events throughout the study affected the results more than expected. Cohort 1 students experienced traveling, season and climate changes, jet lag, and illness during the last week, which coincided with spring break. During that week participants who were tracking productivity were negatively affected as they had fewer reasons to be productive, while people tracking their sleep no longer had a set schedule.

However, since people's lives in general can be busy with activities and events, it is almost impossible to allocate a whole month without expected interruptions. Spring break was also included in the Cohort 2 study because this study had a longer duration (6 weeks instead of 4 weeks). We warned students to find hypotheses that would not be affected by this time off. However, some participants still experienced issues: for example, one participant forgot the charger for her FitBit when she went to Europe for the break and had no data during time period. The longer experiment duration did alleviate some of the effects of those outliers simply by having more measurements.

5.5 Self-experiment analysis method

In Cohort 1, four of the students did not perform any tests to analyze the data—they relied solely on data visualizations for identifying differences between the phases. However, visual analysis is known to be inconsistent and affected by autocorrelation, meaning that it is not a reliable way to reach a valid scientific conclusion [39]. It might be more useful for generating a hypothesis, so we suggested that Cohort 2 students incorporate visualizations in Stage 2, Preliminary Hypothesis Testing.

Cohort 2 students were confused how to interpret the p-values after they performed the required analysis methods. However, the same students reported that the confidence intervals and effect sizes were easier to calculate and understand. Therefore, as previous findings suggest [16], we recommend that those are the methods that novice self-experimenters use if they want to analyze their data.

5.6 Iteration on preliminary hypothesis testing stage

Although the Cohort 2 study allowed a longer self-experiment (6 weeks compared to the 4 weeks in the Cohort 1 study), one participant still considered her statistically insignificant results to be due to an insufficient amount of data. The participant, who explored the relationship between her frequency of exercise and number of initiated conversations on social media, stated that "After conducting this experiment, I strongly believe that the effects of my exercise and physical self-concept is something that would be shaped through months, not a few days or weeks." However, another participant, who studied the relationship between his daily coffee and alcohol consumption and his bodily pH level, reached a conclusive result for both of his hypotheses. The participants' contrasting feelings towards the 6-week timeframe could be attributed to the differences in their results gained from Stage 2, Preliminary Hypothesis Testing. Since the second participant had already tested his method in the preliminary stage and achieved a conclusive result, it was likely that his actual experiment would also be successful. The first participant, on the other hand, altered her tracking method, hypotheses, and dependent variables after Stage 2. Although her altered experimental design might have been more appropriate, she did not have a chance to test them before performing the actual experiment in Stage 3.

Therefore, if at the end of Stage 2, the experimenter decides to drastically alter important pieces of the experiment such as the variables or methods of data collection or analysis, we recommend that she conduct another round of Stage 2 to ensure that her design is still appropriate. If the changes are small and unlikely to affect the overall design, then the experimenter can continue onto Stage 3: Real Experiment.

5.7 Tracking fatigue

Possibly because of the increased study duration, tracking fatigue became a prominent challenge in the Cohort 2 study. All students in the class expressed that they felt decreasing motivation to continue tracking their data. That these self-experiments were part of a class assignment was likely the motivation for them to complete the study. It is interesting to note that students who tracked a variable more than once a day expressed that they wished they had less to track, but they had been too optimistic when they first designed their experiments. We find that self-experiment methods and technologies must address the trade-off between the extending the length of the study to allow for more conclusive results and preventing tracking fatigue.

5.8 Post-Experiment Behavior Change

In an informal poll, only two Cohort 1 students and four Cohort 2 students said they might continue with self-experimentation after the assignment, but in a more passive manner focusing mainly on self-tracking. Given the nature of the study, we do not have data on whether students who found a positive change maintained the new behavior.

An interesting aspect of a self-experiment is its ability to lead to long-term behavior change. Even if users do not continue with self-experimentation, it is important to consider whether self-experimentation tools should also be responsible for helping users maintain the intervention behavior. Perhaps it is not only important that people discover a causal effect, but also that they continue to sustain the new behavior. On the other hand, maybe not all personal informatics tools need to enforce positive behavior change; instead, they could provide as much information as possible to users and let them make an informed decision.

Table 5. Summary of the challenges identified by both cohorts and their suggested mediation.

Challenge	Suggested Mediation
collect reliable self-tracking data	explore various tools
generate a testable hypotheses	explore variables; iterate on hypotheses
support iteration	conduct preliminary hypotheses testing
control for carryover effects	use randomized phase design or randomize each measurement
conduct and interpret statistical analyses	calculate mean differences and size of the effect
avoid tracking fatigue	automate tracking; conduct Bayesian analysis

6 PROPOSED SELF-EXPERIMENT GUIDELINES

6.1 Stages of the Self-Experiment

The cohorts of self-experimenters helped expose the challenges and tensions we described in Section 5. Here, we revise our initial guidelines, and offer them as suggestions on one proposed way of running a self-experiment, meant to empower novices who are looking for guidance.

1) Stage 1: Exploration – try out any devices and variables you think you might be interested in tracking.

2) Stage 2: Preliminary Hypotheses Testing – formulate hypotheses and perform a two week test to assess the data collection, measurement, and analysis methods, and to operationalize the variables.

3) Stage 3: Actual Experiment – either a completely randomized ABAB phase design with a set length, or a Bayesian ABAB phase design without a set length (as discussed below).

6.2 Collecting reliable self-tracking data

As summarized in Table 5, one of the first tensions that self-experimenters need to deal with is between their desire to perform the experiment and their lack of knowledge on how to collect reliable data. One suggestion we have in order to address this issue, is that the self-experiment starts with an exploration of what variables to track and how to track them. This gives the novice a chance to try out various tools and familiarize herself both with what is possible, but also what is useful and reliable.

6.3 Choose testable hypotheses

Another tension we identified is between the need for a testable hypothesis and the lack of clarity on how to come up with one. Based on our findings, we suggest that after the initial Exploration stage, the experimenter picks what she thinks her variables should be, and then designs and conducts a mini self-experiment that runs through the basic structure of the Real Experiment. The goal of this second stage is to operationalize the intervention, variables, and measurements in order to make sure that the intervention is significant enough to make a difference. We discuss how tools can help self-experimenters choose appropriate variables by combining Karkar et al.'s framework with Lee et al.'s use of SMART goals for behavior change in the Discussion section [25, 30].

6.4 Iterate on study design

As we saw in the two cohorts, iteration on both the tools and the design and hypothesis is vital to a successful experiment. Therefore, our suggestion is for experimenters to also visualize their data from this Preliminary Hypothesis Testing stage to help them adjust their initial hypotheses, and all the elements of their designs. Self-experimenters should not begin the real experiment until they feel confident in their design, based on the results of the Exploration and Preliminary Hypothesis Testing phases.

6.5 Control for carryover effects

Carryover effects can play a big part in the results of a self-experiment, as we saw in Cohort 1, so it is an important challenge that we need to address before running the Real Experiment. Here we outline two suggestions on how to account for it and discuss them in detail. For their experimental design, we suggest that self-experimenters either use (1) the ABAB phase design explained above or (2) randomize the condition per trial, depending on the size of the carryover effect. If there is no carryover effect that would influence the dependent variable, a completely randomized design is most appropriate. For example, showering before bed is not expected to have a carryover effect when the dependent variable is the time it takes to fall asleep immediately after. Coffee, on the other hand, has long-lasting effects on the body and might affect time to fall asleep on the next night; thus, it is not recommended to use a completely randomized condition with this independent variable.

In most cases, however, we cannot be certain if there will be a carryover effect. Therefore, based on the exploratory study we completed on the two cohorts, we propose that most self-experimenters follow the ABAB phase design in order to test their hypotheses following Stage 2, the Preliminary Hypothesis Testing stage. Our suggestion for the study design is based on the standards identified by Kratochwill et al., but we use those as guidance, and the specifics can be personalized on a per-case basis [27].

According to those standards, self-experimenters should divide their experiment into four phases and allow for at least 5 measurements per stage. While self-experiments can be conducted with fewer measurements, we agree with Kratochwill et al. that more measurements might be better since, as we saw in the study, sometimes people forget to track, or data is just not recorded properly. Furthermore, if there is a strong carryover effect, the first few measurements of each new phase can be excluded from analysis. We cannot suggest an exact number of days for each phase because some experiments (such as testing the time to fall asleep at night) could take only one measurement per day, while others (such as testing how bloated one feels after eating tomatoes) can be measured multiple times each day. An important part of the ABAB phase design is introducing randomization by randomizing when to switch phases.

6.6 Conduct and interpret statistical analyses

An additional challenge that self-experimenters face is how to analyze data in the best way and how to interpret the results. We turn to literature about the appropriate statistical analysis methods for single case experiments, such as Smith and Duan et al. [10, 39]. In the two cohorts of our study, students were asked to run a t-test, calculate the standardized difference of means, and compute the confidence interval and size of the effect for their data. However, the p-value from the t-test was challenging for novices to interpret. Cohort 2 students claimed in their reports that the difference of means and size of the effects made more sense when they were analyzing their results. Furthermore, if at the end of the long self-experiment, the p-value revealed an inconclusive result, it would be uncertain whether it is due to an insufficient number of measurements or to the actual lack of evidence against the null hypothesis.

Therefore, in order to mediate this issue, we follow the recommendation of Smith [39], who emphasizes that a novice should take a simple approach and look at the difference of means. Then in order to interpret the data, she would look at whether the effect is large in the expected direction. Alternatively, one could analyze the data with a two-sample t-test, with the assumption that each measurement is an independent data point even though the data is autocorrelated since it is from the same person. Duan et al. summarize the most common models used for dealing with autocorrelation in the data, but those models might be too sophisticated for novices to apply on their own [10].

6.7 Bayesian analysis as a way to reduce tracking fatigue

One tension that we identified, and was more pronounced in Cohort 2, was between the necessary minimum length of the study and the experience of tracking fatigue. As a possible mediation, we propose the use of Bayesian analysis, as it could be particularly well-suited for self-experimentation. Bayesian methods have been discussed before in relation to single-case design, but the focus has been on meta-analysis across participants [38]. Alternatively, Jones has focused on a Bayesian analysis using p-values as likelihoods [24].

Bayesian-based experiments can use either the ABAB phase design or the completely randomized design. Schmid and Duan further highlight the usefulness of Bayesian methods when the study design is not fixed, as it provides the opportunity to adapt the design as the study is going on [10]. This makes the Bayesian approach unique by allowing self-experimenters to stop the experiment at any moment and see the probabilistic likelihood of their interventions being effective, and simply having this option could reduce the effects of tracking fatigue. Kay et al. explore the benefits of Bayesian statistics, emphasizing that they lead to more reasonable conclusions for small-n studies, and shift the question towards the strength of the intervention effect rather than a binary "does it work" [26]. Kay et al. also discuss that confidence intervals are often misinterpreted, which makes them less reliable for use with self-experiments. The previously mentioned PREEMPT study also suggests using Bayesian analysis on the N-of-1 trials [4].

A Bayesian multi-armed bandit approach like Thompson sampling [2, 6] could also reduce the common problem of tracking fatigue by having the participant do more of the condition that is more likely to be beneficial. Specifically, the random probability that they are assigned a condition is equal to the posterior probability of that condition being beneficial. This method has been used in applications from website testing [37] to education [42] and increases the amount of benefit to users beyond traditional A/B testing.

However, it is important to note that this kind of analysis might be more challenging than a simple difference-of-means test for novice self-experimenters and can require more initial data. Thus, this method could be implemented in a system for guiding self-experiments that would analyze the collected data and provide a probabilistic result on any day. One issue that arises from this, however, is that because of the nature of the self-experiment, if the experimenter looks at the current result of the experiment, she will become biased and the following measurements might be affected. Therefore, while it is possible to check the current probability on any day, it poses the danger of affecting the later actions and their effects.

Bayesian analysis uses prior probabilities to calculate the posterior probability. In the case of self-experiments, prior probabilities might be helpful as they bring in information from previous self-experiments that might be relevant. Duan et al. discuss in greater detail the issues of the analysis of N-of-1 trials, which are important considerations for self-experiments [10].

7 DISCUSSION

7.1 Technology in Personal Informatics

Even though people have been conducting self-experiments for a long time, digital technology has only recently become a part of these experiments. This experiment showed that digital technology can make data collection easier when used either in conjunction with or instead of manual tracking. It also emphasized the need for an exploratory phase during which experimenters could explore various tracking options, and then a test phase during which they could actually try out some tools in the context of their hypotheses.

Some participants in these cohorts expressed how cumbersome it was to manually track their data and were relieved when they found the appropriate tool to automate the process. For example, P2 in Cohort 1 was tracking the amount of time spent on shopping websites: "The issue being, that if I had to manually record each time I visited a website, I would be conscious of visiting the site and this would thus skew the data. I was able to circumvent this by installing the Chrome plugin TimeStats that silently tracked every site I visited on Chrome,

allowed me to categorize the sites, and computed several statistics for me." Beside the initial set up, P2 did not have to engage with the plugin at all throughout the experiment.

Even with the assistance of digital technology, however, active tracking is error-prone as people have to remember to do it. Some participants used applications that had to be turned on and off for tracking. For example, all the sleep tracking applications need to be started before going to sleep. One participant, who used the Jawbone Up, woke up twice in the middle of the night to find the device switched out of sleep mode and only had a record of the time after turning it back on for those nights. Similarly, a user who was tracking time spent on her computer reported, "It did not reopen automatically when I restarted my computer so I have a day where I got no data."

Sleep tracking applications allow users to do something that they could not before: track sleep parameters like the time it takes to fall asleep and the percentage of the time in bed spent sleeping. Before these applications existed, users could only track the time they went to bed and woke up, along with subjective measures of how long it took them to fall asleep. An important design implication is the need for automated detection of the user going to bed. The same is true for all other applications that require the user to turn them on and off before engaging in an activity: automatic detection prevents user error and mitigates tracking fatigue.

7.2 Designing Tools for Self-Experiments

In the Cohort 1 study, we attempted to address the pitfalls that Choe et al. [7] point out. However, participants still faced challenges with every step of the experimental process, including designing the experiment, collecting data, and analyzing data. We identified some of the main tensions when conducting self-experiments, and we summarized our suggestions on how to mediate them in a list of self-experimentation guidelines which we then tested with Cohort 2. We provide these guidelines as a response to each of the tensions shown in Table 5 so future developers of self-experimentation tools can use them as one way to provide guidance if experimenters were looking for help.

To address the challenge of finding the most reliable and convenient way to track data, it would be helpful if self-experimentation tools could recommend what variables to track and what the most common methods of tracking are. Thus, in order for designers and developers to create effective self-experimentation tools, they need to find a way to help the user explore various tools for collecting data. This might be especially important for populations with lower scientific or technological literacy, who might need more guidance from the moment they decide to start self-tracking. Complete novices might not know what devices and applications to even look at or what variables they might want to track, so a possible first step could be to prepare a guide of most commonly tracked variables and what people used to track them. This process would be a part of the Exploration Stage in our model.

In order to address the challenge of finding an appropriate testable hypothesis, we recommended that self-experimenters explore different variables and iterate on their hypothesis formulation. Similarly to showing a list of common variables and their tracking methods, self-experimentation tools might show a list of hypotheses that were commonly tested for the variables and tracking tools they picked. Furthermore, the self-experiments tool could guide the novice through setting up the self-experiment by asking a series of simple questions. Table 6 shows these sample questions, based on Karkar et al.'s framework [25]. They can be further combined with Lee et al.'s use of SMART (specific, measurable, actionable, realistic, timely) goals to make sure the hypotheses naturally lead to a more successful behavior change [29, 30]. For example, if a user picks "sleep quality" as her dependent variable, and "exercise" as her independent variable, the tool can then ask more specifically what she wants to track and give further options such as "time to fall asleep" and "frequency of exercise."

Furthermore, we need personal informatics tools that, by design, emphasize the iteration on the hypothesis and thus encourage and enable users to perform rigorous self-experiments. The Preliminary Hypothesis Testing

Table 6. Suggested questions for novices to match Karkar et al.'s framework [25]. The answers to all yes/no questions should be "yes." (DV – dependent variable, IV – independent variable).

Karkar et al. Absolute Requirements	Our Suggested Questions
DV: Well-specified (not part of the original Karkar et al. requirements)	These are the most common things people have tried to improve. Pick something you want to improve: [list of most common DVs], or something else. Now, pick a more specific aspect of your DV: [list of specific aspects]
DV: Recurrent episodes or flare-ups	Is your DV something that happens or that you do more than once in a lifetime?
DV: Quantifiable and measurable	Is your DV something that you cannot change just because you want to? Can you measure your (DV) either with a device or by hand?
IV: Controllable and actionable	Can you change your IV just because you want to? Can you measure your IV with a device or manually by hand?
IV: Well-specified	These are the most common things people have tracked. Pick something you think might influence your DV: [list of most common IVs], or something else. Now, pick a more specific aspect of your IV: [list of specific aspects]
DV: Follow the application of the independent variable within a defined period	Does your DV happen after your IV (within a reasonable time)?
IV and DV must not result in any serious health risks (immediate and/or long-term)	Is it safe for you to change your IV and DV (there are no health risks)?
People must be uncertain about the effect of the independent variable on the dependent variable(s)	Do you want to find out how your IV affects your DV?

stage of our model would be the perfect time to do so. Such tools need to guide users through the initial stages of the self-experiment to operationalize their variables (as in Table 6), visualize their preliminary data, and generate hypotheses. By conducting a more scientifically rigorous experiment, users are less likely to be affected by confounding variables and are more likely to reach a conclusive result in a shorter period of time.

A crucial part of any experiment is selecting the appropriate study design. However, as we saw in our study, this was challenging for the novices of Cohort 1, who received no guidance, as their designs were flawed from the beginning. Therefore, this would be an important piece that self-experimentation tools can help with—the designers and developers of such tools could lead the user through a series of questions in order to find the best study design. We summarize those questions in Table 7.

For example, one of the main tensions in our study was between a fully randomized experiment and the possibility of a carryover effect, e.g., if the person chooses to experiment with sleep, the application can easily

Table 7. Suggested tasks to further guide the self-experiment beyond the choice of variables.

Tool side	User side
Choose variables to track	Use questions from Table 6.
Confirm hypothesis question	"Will I fall asleep faster if I exercise for 30 minutes that day?"
Guide user towards most appropriate study design: either completely randomized design or randomized AB phrase design	"Looks like you are tracking your sleep, and it might take a few days for a change to show its effect on sleep. So it's best to follow a "randomized phase design" which means that you will be doing the same thing a few days in a row. We will remind you every day about whether you should exercise tonight or not."
Use Bayesian statistics to analyze the results on the backend of the tool.	Advanced self-experimenters can perform further analyses
Present the results in a manner that novices might be most comfortable with	"If you exercise tonight, there is a 30% chance that you will fall asleep faster."

point that that anything affecting sleep might involve a carryover effect, therefore a randomized phase design might be better than a completely randomized one.

The statistical analysis methods of participants in both cohorts delineated a clear tension between conducting sound analysis of the data and reaching easy to understand findings. This is another a piece of the self-experiment that would benefit greatly from the help of a personal informatics tool. Designers could create template tools for commonly tracked activities. For example, 13 students tracked similar activities, such as sleep quality. Although the students were introduced to various experimental methods, many were still not confident in their skills and their ability to choose the appropriate kind of analysis. Self-experimentation tools could be designed in a way that provides both basic and advanced means of computing statistics after running an experiment: the analysis method could be selected based on the individual's interest and the variables they want to track. One of the simpler methods we suggested based on Smith was to use mean differences and look at the size of the effect [39]. However, there are some more sophisticated common methods in the literature for analyzing data of N-of-1 trials [10]. The tool could present such sophisticated modules and others like intervention analysis [18, 23] and provide further guidance on when to use each tool.

One final tension identified by the study was between the length of the experiment and the needed quantity of data. If a variable can be measured only once a day, the suggested minimum study length, according to the single-case design standards, is twenty days [27]. In the Cohort 2 study, the duration was extended to six weeks as it was a semester-long project. The novice self-experimenters expressed high levels of tracking fatigue by the end of the study. In our suggestions, we recommended, similar to Duan et al., that Bayesian analyses are used in order to mitigate some of those effects, but further work is needed on developing tools that support such methods [10].

Thus, researchers, and perhaps designers of self-experimentation tools in particular, need to further investigate how to strike a balance between this tracking fatigue and the participants' desire to have tracked more variables. We can turn to behavior change literature about possible solutions focused on helping users stay motivated to keep tracking throughout the duration of the experiment. At the same time, technologies need to allow for the

passive tracking of a wider range of variables. One suggested way to mitigate this tension is also to build tools that include methods for calculating the power of the experiment, which would help determine how many data points are needed to find an effect.

7.3 Limitations

One of the main limitations of this study is that it was conducted in a class setting in a university. While we did not specifically answer questions about how to conduct self-experiments beyond what has already been discussed in the paper, it is important to note that students were free to talk amongst themselves and this could have affected how they chose their variables and the rest of the methods. The most important effect of the classroom setting was perhaps that despite the immense tracking fatigue, students continued with the experiment, even though, as they pointed out, they would have given up if they were tracking on their own.

Another limitation of the study is that we conducted this study with two cohorts of university students, who had relatively high statistical and experimental literacy. We have suggested some ways to make self-experiments more accessible to people who lack this kind of background, but we have not yet tested them on such a population.

However, both of these limitations were necessary to help us create a controlled environment where we could present participants with the exact methods we wanted to. Future work could focus on developing a self-contained self-experimentation tool that can allow studies with broader populations outside the class environment while still preserving the ability to control what is being suggested to users.

While we suggest questions in Table 6 and tasks in Table 7 to guide users to perform self-experiments, there might be further limitations that we have not yet addressed. Our goal was to make the questions as straightforward as possible. However, the language and format might need to be altered for specific populations. Veterans, for example, a high percentage of whom suffer from post-traumatic stress disorder (PTSD), might need to specifically focus on self-experiments related to sleep, which is particularly affected by PTSD. Therefore, a self-tracking tool for this vulnerable population would have to be more sensitive towards the kinds of variables they can track related to sleep and interventions that might be most effective for them (such as cognitive behavioral therapy [17]), rather than for the general population. Similarly, the questions and tasks might need to be fine-tuned to address the needs of other specific populations, but the overall framework would remain the same.

8 CONCLUSION

We described a systematic analysis of self-experiments conducted by two cohorts of 34 novices. The first cohort of participants was given minimal guidance, and the lessons we learned from how they conducted and analyzed their experiments were turned into guidelines for the second cohort. We further iterated on these guidelines based on what we observed in the second cohort. We present these guidelines as one way of conducting self-experiments, aimed at novices who want to self-experiment. Based on our observation, our guidelines offer an iterative structure for designing self-experiments, and propose a Bayesian approach to making statistical conclusions as better suited to self-experiments so they can be shortened or extended while ongoing.

Our work contributes to the broader understanding of personal informatics, extending prior work that emphasized the importance of self-experimentation and showed that self-trackers often lack the background to run a rigorous experiment. We learned how people conduct self-experiments when they are given guidance and a basic understanding of experimental design. This allows us to move from broad population studies, which are often cast too widely, to single-case studies which are immediately relevant and targeted to oneself.

ACKNOWLEDGMENTS

The authors would like to thank the students in the seminar classes who participated in the study. This research was supported by NSF grant IIS-1656763.

REFERENCES

[1] 2015. Fitbit. (2015). https://www.fitbit.com/

[2] Shipra Agrawal and Navin Goyal. 2012. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *Proceedings of COLT*. 39–1.

[3] David H. Barlow, Nock K. Matthew, and Michel Hersen. 2008. *Single case experimental designs: Strategies for studying behavior for change*. Pearson.

[4] Colin Barr, Maria Marois, Ida Sim, Christopher H. Schmid, Barth Wilsey, Deborah Ward, Naihua Duan, Ron D. Hays, Joshua Selsky, Joseph Servadio, and others. 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] Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. 2013. Health Mashups: Presenting Statistical Patterns Between Wellbeing Data and Context in Natural Language to Promote Behavior Change. *ACM Trans. Comput.-Hum. Interact.* 20, 5, Article 30 (Nov. 2013), 27 pages. DOI: http://dx.doi.org/10.1145/2503823

[6] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*. 2249–2257.

[7] 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 (CHI '14)*. 1143–1152. DOI: http://dx.doi.org/10.1145/2556288.2557372

[8] 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.

[9] Joan Didion. 1968. On Keeping a Notebook. (1968). Retrieved September 24, 2015 from https://penusa.org/sites/default/files/didion.pdf.

[10] Naihua Duan, Ian Eslick, N Gabler, Heather Kaplan, Richard Kravitz, Eric Larson, Wilson Pace, Christopher Schmid, Ida Sim, and Sunita Vohra. 2014. *Design and Implementation of N-of-1 Trials: A User's Guide*. Agency for Healthcare Research and Quality. www.effectivehealthcare.ahrq.gov/N-1-Trials.cfm

[11] Eugene S. Edgington. 1987. Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology* 34, 4 (1987), 437–442. DOI: http://dx.doi.org/10.1037/0022-0167.34.4.437

[12] Garabed Eknoyan. 1999. Santorio Sanctorius (1561–1636)–founding father of metabolic balance studies. *American journal of nephrology* 19, 2 (1999), 226–233.

[13] 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*. ACM, 731–742.

[14] Deborah Estrin and Ida Sim. 2010. Open mHealth Architecture: An Engine for Health Care Innovation. *Science* 330, 6005 (2010), 759–760. DOI: http://dx.doi.org/10.1126/science.1196187 arXiv:http://www.sciencemag.org/content/330/6005/759.full.pdf

[15] BJ Fogg. 2015. Tiny Habits. (2015). Retrieved September 24, 2015 from http://tinyhabits.com/.

[16] Martin J. Gardner and Douglas G. Altman. 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)* 292, 6522 (1986), 746–750.

[17] Anne Germain, Robin Richardson, Douglas E Moul, Oommen Mammen, Gretchen Haas, Steven D Forman, Noelle Rode, Amy Begley, and Eric A Nofzinger. 2012. Placebo-controlled comparison of prazosin and cognitive-behavioral treatments for sleep disturbances in US Military Veterans. *Journal of psychosomatic research* 72, 2 (2012), 89–96.

[18] Andria Hanbury, Katherine Farley, Carl Thompson, Paul M. Wilson, Duncan Chambers, and Heather Holmes. 2013. Immediate versus sustained effects: interrupted time series analysis of a tailored intervention. *Implementation Science* 8, 1 (2013), 130–147.

[19] Eric B Hekler, Winslow Burleson, and Jisoo Lee. 2013. A DIY Self-Experimentation Toolkit for Behavior Change. In *Personal Informatics in the Wild: Hacking Habits for Health & Happiness at the ACM-CHI Conference*. ACM.

[20] 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. DOI: http://dx.doi.org/10.1016/j.jcbs.2013.10.002

[21] Jim A. Horne and Olov Ostberg. 1976. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology* 4, 2 (1976), 97–110.

[22] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[23] Bradley E. Huitema, Ron Van Houten, and Hana Manal. 2014. Time-series intervention analysis of pedestrian countdown timer effects. *Accident Analysis & Prevention* 72 (2014), 23–31. DOI: http://dx.doi.org/10.1016/j.aap.2014.05.025

[24] W Paul Jones. 2003. Single-case time series with Bayesian analysis: a practitioner's guide.(Methods, Plainly Speaking). *Measurement and evaluation in counseling and development* 36, 1 (2003), 28–40.

[25] 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* (2015), ocv150.

[26] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4521–4532.

[27] 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. DOI: http://dx.doi.org/10.1177/0741932512452794 arXiv:http://rse.sagepub.com/content/34/1/26.full.pdf+html

[28] Quantified Self Labs. 2012. About the Quantified Self. (2012). Retrieved September 08, 2015 from http://quantifiedself.com/about/.

[29] Gary P Latham. 2003. Goal Setting:: A Five-Step Approach to Behavior Change. *Organizational Dynamics* 32, 3 (2003), 309–318.

[30] 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.

[31] 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 (CHI '10)*. 557–566. DOI: http://dx.doi.org/10.1145/1753326.1753409

[32] Allen Neuringer. 1981. Self-experimentation: A call for change. *Behaviorism* 9, 1 (1981), 79–94.

[33] Gemma Phillips, Lambert Felix, Leandro Galli, Vikram Patel, and Philip Edwards. 2010. The effectiveness of M-health technologies for improving health and health services: a systematic review protocol. *BMC Research Notes* 3, 1 (2010), 250. DOI: http://dx.doi.org/10.1186/1756-0500-3-250

[34] Seth Roberts. 2004. Self-experimentation as a source of new ideas: Ten examples about sleep, mood, health, and weight. *Behavioral and Brain Sciences* 27, 2 (2004), 227 – 288.

[35] Seth Roberts. 2010. The unreasonable effectiveness of my self-experimentation. *Medical hypotheses* 75, 6 (2010), 482–489. http://doi.org/10.1016/j.mehy.2010.04.030

[36] Seth Roberts. 2012. The reception of my self-experimentation. *Journal of Business Research* 65, 7 (2012), 1060–1066. DOI: http://dx.doi.org/10.1016/j.jbusres.2011.02.014

[37] Steven L Scott. 2010. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 26, 6 (2010), 639–658.

[38] William R. Shadish, David M. Rindskopf, and Larry V. Hedges. 2008. The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention* 2, 3 (2008), 188–196.

[39] Justin D. Smith. 2012. Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods* 17, 4 (2012), 510–550. DOI: http://dx.doi.org/10.1037/a0029312

[40] John B. Todman and Pat Dugard. 2001. *Single-case and small-n experimental designs: A practical guide to randomization tests*. Psychology Press.

[41] Allen B. Weisse. 2012. Self-experimentation and its role in medical research. *Texas Heart Institute Journal* 39, 1 (2012), 51–54.

[42] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z Gajos, Walter S Lasecki, and Neil Heffernan. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 379–388.

Received February 2017; revised May 2017; accepted July 2017