A Cohort of Self-Experimenters: Lessons Learned from N=1 Personal Informatics Experiments

Nediyana Daskalova
Computer Science
Brown University
nediyana@cs.brown.edu

Karthik Desingh
Computer Science
Brown University
karthik_desingh@brown.edu

Jin Young Kim
Microsoft
jink@microsoft.com

Lixiang Zhang
Computer Science
Brown University
lixiang_zhang@brown.edu

Alexandra Papoutsaki
Computer Science
Brown University
alexpap@cs.brown.edu

Jeff Huang
Computer Science
Brown University
conference@jeffhuang.com

ABSTRACT

Personal informatics, the trend in collecting and analyzing data about one's own self, has been heavily influenced by the involvement of technology in our daily lives. In this paper we aim to understand what happens when novice self-experimenters are provided with a structured lesson in experimental design. We conducted a month-long study on self-experimentation, where twenty students in a seminar performed a self-experiment of their choice. The students were prepared with background readings and lessons on statistical analysis and experimental design. Their experiments were designed in a structured manner: a specified number of variables to track and a set duration for the study. We also analyze videos of self-experimenters from the Quantified Self community, and compare them to the methods and outcomes used by the students. We find pitfalls that both students and self-motivated *Quantified-Selfers* experience, such as a too short duration of the study and insufficient planning of the set up of the experiment. Based on these findings, we propose an iterative self-experiment design method that addresses those pitfalls. We also discuss broader implications for future self-experimenters and designers of tools for self-experimentation.

Author Keywords

personal informatics; self-tracking; quantified self

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

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

Personal informatics is data about you and for you! It is an increasingly popular trend in collecting, analyzing, and reflecting on various facets of one's personally relevant 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 [17, 4]. The findings from these investigations have shown that people generally perform descriptive analyses, and that tracking one's own behavior is a beneficial process. It is reported that people perform self-tracking to be mindful of their behavior, to motivate themselves for goal achievement, and to improve themselves and their lives in general. Underlying many of these goals, there is a pressing need for self-knowledge. Namely, one needs to understand causal relationships in one's life to achieve these goals.

The current paradigm in research on behavior change as performed in fields such as public health, social sciences, and research initiatives like mHealth [9, 20], is to find generalizable effects about people that can be disseminated to the public. However, by definition there only has to be a small effect on a subset of that population for those 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 [14]. 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. Thus a personalized approach to experimentation is more relevant. The goal of personal informatics is not to discover knowledge about a broad population, but instead to help people learn more about what affects them, and specifically only for the parts of their lives that matter to them the most.

Submitted for review.

Beyond passive monitoring, the next logical step towards an even better understanding of one's self is to actually perform self-experiments, that is to create and test hypotheses on the effect of small behavior changes [11]. However, many individuals that perform self-tracking do not have the capability to conduct analyses or run rigorous experiments, as it is not clear what would be an ideal procedure that they should follow. In this paper we seek to answer the following question: "what happens when people are able to run experiments, perform analyses, and create visualizations by being offered a structured lesson in experimental design?" In this paper we present the findings of a study on self-experimentation, where a cohort of twenty students in a Human-Computer Interaction seminar, each performed an experiment of their choice on themselves as part of a month-long assignment. The students designed hypotheses and tracked the appropriate variables that would be suitable for testing them. They also submitted detailed reports comprising their procedures, a day-by-day journal (both textual and numeric), visualizations, and analyses. While these students were given guidance, they were relatively inexperienced in self-experimentation and were faced with constraints such as the duration of the experiment. To complement findings from the N=1 experiments above, we also survey methodologies from *Quantified-Selfers* to learn about self-experimentation that happens outside the classroom. Following the approaches from Choe et al. [4], we study videos and articles of self-experimentation posted on `www.quantifiedself.com` over the past thirteen months, focusing on experimentation methodology and outcome. This enhances the students' experiments by providing insights from self-tracking enthusiasts who are self-motivated and have no timing constraints. We find that many Quantified-Selfers experimented over longer periods, and their choice of data collection and analysis is correlated with their background in science and engineering. However, they often employed naturalistic data collection and simple visualizations, which seems to suggest a somewhat different nature of self-experimentation.

The main contribution of this work is a series of lessons about self-experimentation from both a class study and an analysis of Quantified-Selfers' videos. Self-experimentation itself is a challenge because it is neither clear what the appropriate procedure for such experiments is, nor is there a body of successful (or unsuccessful) trials to draw from. We believe this is the first meta-analysis of multiple N=1 experiments conducted in a structured environment, where participants are given the freedom to choose their experiment and are provided with some guidance on designing these experiments. We find that: 1) one month is typically not enough to reach a valid conclusion of the self-experiment, 2) an exploratory stage is at the beginning of the study is highly desirable, and 3) iterations on the design of the experiment help determine the optimal combination of tracking tools and variables. We combine these 3 key findings and propose an iterative self-experiment design which can help both future self-experimenters and designers of tools for self-experimentation.

RELATED WORK

Self-Experimentation

Self-experimentation is a type of scientific experiment in which the experimenter herself is the only subject involved [25]. One of the earliest documented examples of this type of research is that of Sanctorius of Padua. In the early 17th century, Sanctorius weighed himself daily over a period of thirty years along with his food and liquid intake and body excretions. This self-experiment led to the discovery of metabolism [8]. In our study some students also logged their weight with electronic devices, which made data analysis later much easier.

Another well-known self-experiment on sleep (a common variable tracked by many students in our experiment) is that of Michel Siffre, an underground explorer who in 1962 spent two months in an underground cave [10]. He lived in darkness with no way to tell the time and would call his team whenever he woke up, went to bed, and had a meal. His experiment started the field of human chronobiology which studies the human circadian rhythm.

Self-experimentation has a long documented history in a variety of research fields such as medicine and psychology. Self-experiments take various forms, with some being momentary while others being longer-term [22]. One example of a brief self-experiment was the notable case of researcher Barry Marshall, who in 1984 drank a petri-disk of Helicobacteria pylori from a patient and soon developed gastritis and nausea, establishing a causal relationship between the microbe and the disease [19]. In this paper we present long self-experiments that lasted a month for each student and involved tracking many variables.

Personal Informatics

People have been tracking data about 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 though we would never forget" [6]. Recently, technological advances have brought self-monitoring to current times, allowing almost anything to be tracked. People can now record not only various aspects of their health such as calories, miles run, amount of time slept, but also how they spend their time and money, and how productive they are. The most common technological tools that assist them in self-tracking are their smartphones and other portable and wearable devices, like the popular FitBit. However, the amount of data people collect about themselves is so overwhelming that specific innovations focus on synthesizing the information from multiple platforms and presenting it to users in a simpler, more understandable form [2].

Systems which help people collect information about themselves are called personal informatics systems. While many people use them to improve their well-being, these systems can also provide general help by enabling personal reflection, self-knowledge, and discovery [17]. The advantage of personal informatics systems is that they have a capacity for collecting and storing data that far exceeds the human memory and for presenting them in different forms. According to studies, short term memory has a limit of how many units it can keep (three to four) [5]. In addition to that, the modality effect in learning refers to the idea that information is more easily acquired when it is perceived in a variety of modes: not only visual through reading, but maybe also auditory or through writing it out [18]. Therefore, a written record of useful information has the advantage of leading to insightful reflection that cannot be achieved through our otherwise limited memory since the question users seek to answer is what affects them in the long-term.

"Quantified Self" is a community of people who use and design tools for personal informatics [16]. Quantified-Selfers, as they are commonly known, hold Meetups around the world, during which people can present what they tracked, how they tracked it, and what they learned from it. Their reports are posted as videos on their website `www.quantifiedself.com`. 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 [4]. They found that Quantified-Selfers compromised the validity of their results and identified 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 what happens when these common pitfalls are sidestepped and focus on the lessons about self-experimentation, where users are given some guidance while designing their experiments.

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 [24]. 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 one [17]. They emphasize that tools for personal informatics should allow users to iterate on the stages of their experiments, which further supports the iterative self-experiment design that suggests that self-experiments themselves should be iterated on in order to find the optimal design.

There are some forms of personal informatics that 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 archive various aspects of their lives as a new way of journaling and reflection [4]. However, in the vast majority of cases, the goal of personal informatics is to understand some aspect of one's self and life, and self-experimentation is the only way to achieve this rigorously. Even outside the Quantified Self community, N=1 clinical trials have the possibility to help individual patients, and large pharmaceutical companies are seeing their potential for future research [3].

Hekler et al. point that many of the current technologies do not provide users with the tools to self-experiment, as knowledge on its own is not enough for behavior changes [13]. As they explain, the key towards forming new habits is to provide a context and link new behaviors to existing ones. Fogg emphasizes in his "Three Tiny Habits System" the importance of fostering new habits through starting with tiny incremental steps based on established behavioral routines [11]. Therefore, when people self-experiment and try to change something about themselves, the knowledge of the change is not enough to turn it into a behavior change. However, they can more easily create a new habit as they are already seeing positive changes from their self-experiment.

Roberts has played a pioneering role in introducing the possibility of self-experimentation to the self-trackers that are new in the Quantified Self community [22, 23]. He ran numerous experiments over a period of twelve years, identifying several novel casual 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 reported to be effective anecdotally. He argues that self-experimentation has several benefits over conventional research, including strong self-motivation, no limit in experiment duration, and easier idea generation and validation because the experimenter becomes the subject himself.

Researchers such as Choe et al. [4] have compiled tips for novice self-trackers based on advice given by experienced ones. In this paper we describe the lessons that we learned from a meta-analysis of multiple N=1 experiments performed in a structured environment where experimenters are provided with the above tips. In addition, we complement the findings from the students' experiments by surveying the methodologies of self-experimentation that can be found in the QuantifiedSelf website.

STRUCTURE OF THE N=1 EXPERIMENTS

We distributed an assignment in a Human-Computer Interaction seminar and had a cohort of twenty undergraduate and graduate Computer Science students run a month-long self-experiment. While there were twenty-one students in the class, there was one who did not consent to disclose their collected data and completed surveys, and thus we omit them from the background examination of the experimenters. In total, the assignment lasted five weeks, with a combined 1 week for planning and analysis, and four 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 their choices.

	Yes	No
Statistics experience	13	7
Self-tracking experience	5	15

Table 1. Number of students with statistical background or self-tracking experience prior to the start of the class

Students were encouraged to use technological tools to assist them in monitoring their variables. Most students used smartphones and wearable devices, either their own or from a loan pool of fitness trackers that we provided, whereas some logged their observations and measurements on spreadsheets. At the end of the month-long experiment each student submitted a report that described their hypotheses, variables they tracked, statistics they used, a day-by-day journal (both textual and numeric), and visualizations and analyses they performed to test their hypothesis. The assignment also asked students to include a discussion of the lessons learned and whether the results matched their expectations.

The class consisted of twelve male and nine female students, representing both undergraduate and graduate students. Every member of the class was given the same set of directions. These directions instructed that they could track any combination of independent and dependent variables, as long as there was a hypothesis that could be tested and that the data regarding the variables could be analyzed. We observe commonalities in the various aspects of the process across all students including variables, confounding factors, and statistical results.

The students had varied experience in experimental design. Upon the completion of class, all students were asked to complete a survey asking them for more information on their level of expertise with statistics and personal informatics. As can be seen in Table 1, most students had statistical background before taking this class, however the majority of them lacked previous experience with self-tracking. In order to ensure that all students had a foundation on personal informatics, a series of basic understanding of statistical tests and knowledge related to personal informatics were disseminated through paper readings and discussions on experimental methods and behavioral analysis prior to the study.

This is a specific population who is capable of quickly learning scientific methods and generating visualizations and analyses for their experiments.

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 this study. Table 2 shows the list of independent and dependent variables chosen by the participants, along with their hypotheses and experimental outcomes. We refer to the participant IDs throughout the paper to discuss specifics of certain experiments.

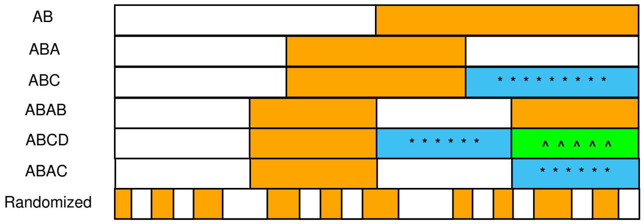

Figure 1. Different methodologies and their time series patterns

It is evident from Table 2 that there is a diverse list of dependent variables that were tracked across all participants, 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. Participants also used different experimental designs and methods of tracking and measuring the variables as per their convenience and available resources. Across the 20 participants, 16 used some form of AB* testing 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 randomization.

Dependent variables can be divided into two categories based on how they are affected by a particular independent variable. One might change gradually and the other one immediately. For example, mood (which is subjectively measured) might be a dependent variable that gets impacted immediately and sleep quality might be something that changes gradually.

APPROPRIATE EXPERIMENTAL DESIGN

According to Choe et al. [4], one of the common pitfalls when conducting self-experiments is the lack of scientific rigor. For this study, we addressed that pitfall by introducing participants to a variety of techniques for conducting and analyzing an experiment. Students read papers throughout the course of the class and critiqued the methods used in them. Thus, they were better prepared to design their own experiments. For example, nearly all students performed a variation of AB* testing, and one used a coin toss as a source of randomization.

Intervention based Interrupted Time Series

In this 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. This change in behavior signifies a new phase of the experiment. Figure 1 shows the different design methods and their phases in a time series experiment. 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. AB* denotes the pattern of behavior A followed by B and then any other pattern later on.

Consider an ideal case where a participant is analyzing one dependent and one independent variable as part of their experiment. We cannot say with certainty that a single design is the best pattern to use, as the design

P	Independent Variable(s)	Dependent Variable(s)	Hypotheses	Design	Result
1	exercised for 30 minutes	sleep quality	more exercise, better sleep quality	AB	YES
	number of steps	sleep quality	more steps, better sleep quality		IC
	type of task	heart rate	heart rate differs between tasks		YES
2	number of classes attended	productivity	less attendance, more productivity	ABA	IC
		time spent online shopping	less attendance, less time spent online shopping		YES
		unnecessary spending	less attendance, less unnecessary spending		YES
3	weather conditions	exercise frequency	better weather conditions, more frequent exercise	Natur	IC
		exercise duration	better weather conditions, longer exercise		IC
4	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*
5	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		IC
		mood	more tea, better mood		IC
6	showered before bed	time to fall asleep	shower before bed, less time to fall asleep	Rand	IC
		resting heart rate	shower before bed, lower resting heart rate		IC
		amount of restful sleep	shower before bed, more restful sleep		IC
7	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
8	exercised, took supplements	weight	more exercise & supplements, more weight	AB	YES*
		heart rate	more exercise & supplements, higher heart rate		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*
9	electronics used past 9pm	sleep quality	no electronics after 9pm, better sleep quality	ABA	IC
		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*
10	ran for 30 mins	sleep quality	running affects sleep quality (significantly)	AB	IC
		heart rate upon waking up	running affects heart rate (significantly)		IC
		sleep quality	running and sleep are not independent		YES
		steps in bed	running affects steps in bed (significantly)		IC
11	minutes being tickled	amount spent per week	tickling will increase money spending	ABCD	IC
		morning mood	tickling will improve mood		IC
		sleep quality	tickling will improve sleep quality		IC
		weight	tickling will increase weight		IC
12	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		IC
		number of awakenings	drinking apple cider vinegar, better sleep quality		IC
		time to fall asleep	drinking apple cider vinegar, better sleep quality		IC
13	amount of coffee consumed	productivity	more coffee, improved productivity	ABCD	IC
		sleep	more coffee, less sleep		IC
14	ran in the morning	heart rate	morning runs reconcile midday and morning heart rates	ABAB	IC
		heart rate	morning runs reconcile midday and evening heart rates		IC
		daily PSS score (stress)	leads to lower total PSS score		IC
15	amount of screen time per day	sleep quality	looking at a computer at bed time, poorer sleep	Natur	IC
	screen-less time before bed	sleep quality	the lower the temp, the more		IC
16	mean daily temperature	hot beverage drank in the day	the lower the temp, the more hot beverages drank	Natur	YES
		self-report feeling of laziness	the lower the temp, the lazier about working		IC
17	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*
18	used time blocking	mood	time blocking will improve mood	AB	IC
		sleep quality	time blocking will improve sleep quality		IC
		productivity	time blocking will improve productivity		YES
19	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
20	consistency of bed/wake time	sleep quality	fixed sleep time window, better sleep quality	ABAC	IC
		productivity	fixed sleep time window, increase productivity		IC
		reduce tiredness	reduced tiredness levels		YES

Table 2. Personal Informatics Experiments (* visual analysis, IC - inconclusive, Natur - Naturalistic, Rand - Randomization)

Design	Accepted		Rejected		Inconclusive	Total
	Stat	Vis	Stat	Vis		
AB	4	1	0	2	5	10
ABA	4	3	1	0	2	10
ABAB	2	0	0	0	3	5
ABAC	1	0	1	0	4	6
ABCD	0	0	0	0	3	3
Rand	0	0	0	0	3	3
Naturalistic	1	0	0	0	3	4

Table 3. Summary of hypotheses' results from different experimental designs chosen by the participants (Stat = analyzed statistically, Vis = analyzed visually)

methodology is subjective and heavily dependent on the variables chosen. For example, if the independent variable is "Running 5 miles" and the dependent "Mood" is being analyzed after the run, then the AB* pattern might not be appropriate, and a better design might be a Randomization experiment which is discussed in the next section. But if the participant is trying to measure their "Sleep quality" when they run during the day, then they might have to experiment with this for a while in order to know how it is affected. Thus, we cannot decide which kind, if any, of the AB* pattern is best without knowing whether the dependent variable is immediately or gradually affected by the independent variable and what the duration of the experiment should be.

In this four week experiment, the participants who chose the AB* method of experimental design allotted at the minimum of one week for each of the phases. Now we discuss the results obtained based on the design methods chosen by the participants. In Table 3 the results are consolidated with respect to these design patterns. Patterns AB and ABA are predominantly chosen by participants for their experiments. One example of a participant using the ABAC pattern is P5 who used it for analyzing the effect of green tea consumption on sleep quality. In the C phase, P5 chose a different consumption strategy to see if drinking tea before and after meals had any significant change. Another example is P13, who chose ABCD for analyzing the effect of coffee consumption along with its amount on the dependent variables. Thus, P5 and P13 chose to vary their independent variables and test the impact those changes had on their dependent variables. Therefore, if there are numerous variations in the independent variables in an experiment, design patterns like ABAC and ABCD can be used. At the same time, the duration of the total experiment should be sufficiently large to explore these variations. P13's results were inconclusive at the end of the experiment and this is partially due to the fact that there were not enough data points.

Randomized conditions

According to single-subject experimental design literature [7], randomized single-subject experiments can be helpful for individualized treatments of patients and systematic replication can lead to insights about a larger population. These experiments use randomized tests to assess the efficacy of a treatment. The tests are called "randomized" because they are based on the random assignment of treatment. A major concern among researchers who oppose randomized single-subject experiments is that the treatment might harm the subject if administered randomly [7]. However, for self-experiments with personal informatics, which some people call "soft science" [21], there is no such danger. Further, since the experimenters are also the subjects in this case, they can always stop the experiment if they feel any discomfort. Only P6 in the our study chose to use randomized conditions as their design methodology. P6's independent variable was whether or not she showered before going to bed at night, and she decided that 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." Therefore, randomized conditions are 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.

Level of confidence	1	2	3	4	5
Number of participants	0	2	7	8	3

Table 4. Participants' confidence levels about their experiments from 1 (least) to 5 (most).

CHALLENGES IN MEASUREMENT AND TRACKING

One of the main challenges in our study was that sometimes the dependent variables were actually manipulable by the experimenter. Some students were not careful enough in choosing their independent and dependent variables at the beginning of the study, so their results might have been affected by confounding factors. Furthermore, many students tracked productivity, but there is no set standard metric for measuring or tracking it, so these results were the most subjective ones. At the end of the study, when a survey was sent out with a question "How confident were you about the experiment performed? (Were the variables tracked reliably)" with scale of 1–5 from least to most confident, the most common responses were in the range 3–4 as shown in the Table 4.

Another challenge was that students ran the experiment for only four weeks, with the last two being the week with midterms and then spring break. A longer duration of the experiment might have eliminated or at least alleviated the effect of these changes in their schedules. The short duration of the study also led to an insufficient number of data points, which hindered the observation of statistically significant relationships between variables. Many students reported that if they were to do the study again, they would extend the duration of the experiments.

A better study might have lasted for at least two months in order to have a month for each phase of the AB*.

Furthermore, life events throughout the duration of the study affected the results more than expected. Since the last week of the experiments was during spring break, students experienced traveling, season and climate changes, jet lag, illness, and so on. Participants tracking productivity were negatively affected as they had fewer reasons to be productive with midterms, homework, and projects out of the way during spring break, whereas people tracking their sleep no longer had a set schedule to stick to.

Another challenge in measurement and tracking was that some variables actually vary throughout the day. For example weight fluctuates throughout the day, thus the choice of time of the day one measures their weight would affect their findings. As a possible solution, participants could record the variable multiple times per day in order to visualize the variation of data.

CHALLENGES IN ANALYSIS

Based on the findings from Choe et al.'s paper [4], participants need some background in analysis and visualizations in order to effectively design their own experiments and learn from the results. Students in this study built the essential background by reading research papers that described analysis methods and visualizations and through in-class discussions. They were also introduced to topics in experimental methods and behavior analysis. However, building on Choe et al.'s advice, we learn that it is not enough to have a basic understanding of methods of analysis, but rather there is a need for specific useful types of analysis that are relevant to single-subject experiments.

As noted earlier, 15 out of 20 students were fairly confident about the analysis they conducted. From the 2 participants with lowest confidence, one (P11) used only a correlation coefficient to analyze his data, and the other (P6) used correlation, t-test, and Hedges' g. From the 3 students who were absolute confident: one (P4) used only visualizations and calculations of the mean and variance to perform analysis of his data, another one P(9) used visualizations and Hedge's g, and the third one (P7) used correlations, calculations of mean, standard deviation, and variance, Hedge's g for effect size, and chi-squared tests to find p-values.

Single-subject experiments require careful analysis, as the data is from within the same subject, and it might not be normally distributed. Four of the students did not perform any tests to analyze the data—they relied solely on visualizations.

SURVEY OF QUANTIFIED-SELFERS' EXPERIMENTS

So far we discussed the N=1 experiments from students in the course. These results illuminate the possibilities of what can be learned by relatively inexperienced self-experimenters when proper guidance is provided. On the other hand, there are burgeoning communities of self-trackers such as Quantified Self, and many of these people perform self-motivated experiments without any constraints such as the duration of experiments.

Since the experiments done by these enthusiasts should provide additional lessons on self-experimentation, we present a survey of self-experiments from recent reports found in Quantified Self. Note that these people are early adopters of self-tracking whose characteristics are different from general public, as is evident by the distribution of backgrounds detailed below. Also, their presentations of self-experiments are likely to suffer from reporting bias common in academic literature, given that negative results are likely deemphasized or even omitted.

Methodology

For our survey, we collected 65 reports of self-tracking. These include videos, slides, and articles that are found in or linked from QuantifiedSelf.com between August 1, 2014 and August 31, 2015. Since our goal was to focus on self-experimentation, we used only those with (1) specific independent and dependent variables, and (2) the descriptions of data collection and analysis results. This left us with 20 experiments out of 65 reports we collected. When a report describes multiple experiments, we used only the most recent one. For each experiment, we identified the following variables: the background of the subject in Science, Technology, Engineering, and Mathematics (STEM), the duration of the experiment, independent and dependent variables, experimental design, analysis method, and the outcome of the experiment.

Findings

Table 5 describes the summary of the experiments we collected. The subjects are numbered by the order in which the experiment was reported on the Quantified Self, with the latest experiment on the top. We identify individual subjects as Qn where n is the subject's ID. Reports often missed some details of the experiment and were left as a blank. Let's first look at the background of experimenters. Many subjects were from science or engineering (10/20), others were suffering from symptoms they were tracking (Q18), and then there were CEOs of tech startups (Q11, Q14). Everyone had sufficient reasons to learn experimental methods.

In terms of the duration, except for two experiments that lasted for about a month, all experiments were run for a median duration of a year. This contrasts with the duration of students' experiments (four weeks), stressing the importance of sufficient time in self-experimentation. Also, many of the subjects report that they have been running multiple experiments on the same or related topics, whereas students did not have chances to correct or re-run experiments once they realized what they could improve. This leads us to one of our key findings that self-experimentation is an iterative process which benefits from the accumulation of experience over time and makes it an integral part of the iterative self-experiment.

The domain of experiments covered mostly were health, mood, or cognitive status, except for one subject (Q6) who tracked the response rate in online dating. This shows similar distribution to the students' experiments where health and mood were dominant topics. The majority of subjects (16/20) used multiple independent variables, and 30% (6/20) of subjects used multiple dependent variables. This contrasts with students' experiments where the majority used a single independent variable. This might be related to the longer duration of the experiments in overall, which would have enabled them to observe the effect of multiple independent variables.

As for experimental design, 10/20 subjects collected data and introduced some form of intervention, while the rest recorded data in naturalistic settings. The backgrounds of these participants seem to have slightly affected their choice of experimental methodology. Among 10 subjects with STEM backgrounds, 6 used controlled experiments, whereas the other 4 chose to use naturalistic study. While the difference is not significant, training in STEM may have taught them the importance of experimental control in drawing causal relationships from data.

In terms of analysis methods, only three subjects reported the results of null hypothesis statistical significance test (NHSST), whereas the rest used visualization and summary statistics (i.e., correlation coefficient) to draw conclusions. Again, all three subjects who used NHSST had a STEM background. In terms of results, everyone reported some causal relationship, perhaps reflecting the reporting bias as is commonly found in published results. However, many of reported results could hardly be considered "publishable" according to the standard of conventional research methodology. Many of the "experiments" resemble exploratory data analyses, with reports of correlations between variables based on naturalistic observations, rather than rigorous controlled experiments.

So what can explain this seemingly "sloppy" standard of rigor? A possible explanation is that many of these participants did not get proper training in experimental design and statistical analysis. Further, Quantified Self is not peer-reviewed. We believe that there are several factors that make self-experiments different from traditional experiments. First, doing a controlled experiment on yourself is not feasible in some cases, or takes considerable effort in following the experimental conditions. Second, for many of these experiments, experimenters' subjective feelings would be strong indicators of the success. While these experiments could still benefit from rigorous statistical analyses, we believe that the standard for determining what is conclusive for self-experiments may be different from the one for traditional experiments.

DISCUSSION

Post-Experiment Behavior Change

An important aspect of a self-experiment is if it can lead to a behavior change based on the findings or not. In an informal poll, only two students in the class said they would continue with self-experimentation after the assignment. One revelation from this is that it is not only important that people discover a causal effect, but they should also wish to act on it. But perhaps not all personal informatics tools need to enforce positive behavior change, but instead they should provide as much information to the user as possible, and let the user make an informed decision.

Technology in Personal Informatics

Even though people have been doing self-experiments for a long time, technology has only recently become a part of it. Manual tracking is still used, but based on the results from this experiment, technology makes the process easier and allows for data to be collected seamlessly. However, whenever manual tracking is involved with technology, there are still issues with recording data as people have to remember to do it. Out of the 11 students that used a device that required them to manually start and stop it, 2 had at least once a problem with the data collection.

Some participants in this study expressed how cumbersome it would be to manually track the data and were relieved when they found the appropriate tool to automate the process. For example, P2 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." Besides the initial set up, this user did not have to engage with the plugin at all throughout the experiment in order to keep tracking the data.

Other participants, however, used applications that required them to turn them on and off when tracking. For example, all the sleep tracking applications need to be started before going to sleep. One participant, who was using Jawbone Up, woke up twice in the middle of the night to find the device switched out of sleep mode, so he only had a record of the time after turning it back on for those nights. A similar issue was encountered by the user who was tracking her time spent on her computer—"It did not reopen automatically when I restarted my computer so I have a day where I got no data."

Sleep tracking applications are already allowing users to do something that they could not before—track sleep parameters like time it takes you to fall asleep, percentage of the time in bed spent sleeping, etc. Before their existence, all users could manually track was the time they went to bed and woke, and 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.

Experience Sampling as an Alternative Tracking Method

ID	STEM Background	Duration	Independent Variables	Dependent Variables	Exp. Design	Method
1	Scientist	Years	Drug / Lifestyle / Mental	Sleepwalking	AB	Vis+NHSST
2	Engineer	Years	Sleep time / hour	Quality of Day	Naturalistic	Vis+NHSST
3	M.S.	240 days	Amount of soy in diet	Reaction time	ABCA	Vis+Stat
4	Ex-Engineer	120 days	Fish oil intake	Reaction time	Randomization	Vis+NHSST
5		27 days	Heart Rate Variability	State of Flow	Naturalistic	Vis
6		3 years	Message length, wording, timing	Response rate	Naturalistic	Vis+Stat
7		28 years	Exercise / Diet	Weight	Naturalistic	
8		130 days	Dinner time / food / Sleep time	Sleep quality	Naturalistic	Vis+Stat
9	Physical Therapist	160 days	Diet (seed / fiber / water)	Stool quantity / quality	ABACAD	Vis
10	Engineer		Mental / physical / activity / ...	Mood	Naturalistic	
11		200 days	Cycling (with different intensity)	Blood pressure	ABC	Vis
12		330 days	Food composition	Allergy (blurping)	ABACAD	Vis+Stat
13		2 years	Alcohol / Sleep time	Sleep quality	AB	Vis
14			Exercise / Activity / Diet	Weight / Sleep quality	Naturalistic	Vis
15			Exercise /Diet	Weight	Naturalistic	Vis
16	Physician	7 years	Lifestyle change	Weight / Activity / Strength	AB	Vis
17	Professor	7 months	Diet / Drug	Neuro-cognitive function	ABC	Vis
18		Years	Diet / Sleep	Parkinson's Disease symptom	ABC	Vis
19	Data Scientist	340 days	Time / DoW / Season	Sleep quality	Naturalistic	Vis
20	Patient / Chemist	30 years	Life events / Food	Diabetes symptoms	Naturalistic	

Table 5. Survey of Quantified-Selfers' self-experiments; NHSST is the null hypothesis statistical significance test

Experience sampling is a method for collecting information from participants as samples of their behavior and experiences. In our experiments, only one student (P18) used this method to track his mood and productivity. He used an app called Reporter to design his own questions for sampling, and receive randomly timed reminders, which could reduce bias. Besides him, 6 other students tracked productivity and 2 others tracked mood. However, none of them used experience sampling: for example, they recorded manually how productive they were each hour of the day. However, as one user put it, he would often enter "data for several hours at later times (e.g. lunch and coffee break)," which might have affected his productivity ratings. Thus, if they had iterated on the design of their experiment, they could have started using apps like Reporter to improve the quality of the collected data.

Short Length of Study Leads to Inconclusive Results

If an experimenter did not manage to disprove their null hypothesis, then the results of the experiment were deemed inconclusive. There are various reasons why it was not possible to reject the null hypothesis, including the already mentioned one that the length of the study might have been too short, so there simply were not enough data points. Another reason is that the experimenter did not design their experiment appropriately and might have picked improper variables.

Inconclusive results do not mean that the hypothesis was incorrect, but rather that the null hypothesis was a plausible outcome due to chance. In other words, the statistical test performed did not pass the significance threshold needed to reject the null hypothesis, which means that there was no strong evidence against the null. Thus, either the null hypothesis is actually true and there really is no significant difference, or it is false, but there is not enough data to prove it. Therefore, a way to address this issue is to extend the length of the experiment.

The inconclusive results in AB* experiments could also be caused by behavioral effects that carry over across different phases. ABA theoretically mitigates the effects of the participant behavior before the start of the experiments. However, variables such as sleep quality might take longer to be affected by changes. For example, a common sleep hygiene guideline is to set a consistent bed and wake times because over time your body gets used to it and it is easier to wake up in the morning [1]. A change like that cannot happen overnight.

Students in this study were limited by the time frame of the assignment, so they could not continue their experiments. However, many of them were aware of the consequences of this limitation and noted that their study might have had conclusive results if it had been longer. Thus, an important design consideration would be to teach experimenters about the importance of a prolonged study, and that being unable to reject the null hypothesis does not necessarily mean that their hypotheses were wrong. Experimenters could also be nudged and motivated to continue with their experimental set up to collect more data, which might lead to conclusive results.

How Students Differed from Quantified-Selfers

Based on the results from both the in-class study and the analysis of the Quantified-Selfers' videos, we propose the an iterative self-experiment design, which is comprised of 3 key components. One main finding from the in-class study was that regardless of what experiment and analysis students chose, 4 weeks was prevalently too short of a time-frame as power decreased since days had to be divided across the different conditions. In contrast, when Quantified-Selfers conducted self-experiments, they ran them for months or even years. In addition to that, personal events and other regular schedule disruptions spring break led to inconclusive results. Thus, the optimal design would last longer, increasing the power of statistical tests and allowing for more conclusive results.

An important lesson from the Quantified-Selfers is that they usually iterate on the design of their experiments. When the method of data collection or the variables turn out to be unsuitable for their experiment, they change them. Thus, for the design of the perfect experiment, they go through a series of iterations until they find the optimal combination of tracking tools and variables they want to track. The students in the class study could not iterate on their design. Due to the constraints in the length of the experiment, if they had tried switching their variables or even the tool they were tracking them with, they would have had even less data so they would not been able to analyze it at all. Thus, the second part of the iterative self-experiment design focuses on iterating on the design of the experiment before the actual data collection begins in order to avoid errors in data collection methodologies and even picking inappropriate variables.

The third component of the iterative self-experiment design is an exploratory stage at the beginning of the experiment. During this stage, people should be encouraged to try out different tracking methods and variables, and iterate on the process until they find the best design that works for their lifestyle and variables they want to track. We find that people who used the naturalistic design throughout their whole experiment, usually used visualizations and correlations to analyze their data. While this may be a valid test as we mentioned previously depending on the type of data they are collecting, one cannot prove causation without having a controlled experiment. Thus, if a person finds a high correlation between variables in the the naturalistic exploratory stage, the best way to proceed may be to design a controlled experiment where they can actually test for a causal relationship.

Designing Tools for Self-Experiments

In our study we addressed the pitfalls that Choe et al. [4] point out. However, participants still faced challenges with every step of the experimental process: setting up the experiment, collecting the data, and analyzing them. Therefore, we are providing future developers of tools for self-experimentation some additional insights.

Although the students had been introduced to various methods of conducting experiments and analyzing them, many were still not confident in their skills and whether they conducted the appropriate kind of analysis. Hence, there are many ways in which designers can create tools to ease the entire process. One set of tools could focus on the statistical aspect of the experimental analysis—by making it easier for experimenters to compute statistics after running an interrupted time series or randomized experiment. It could even present more sophisticated tools like intervention analysis [12, 15], and provide further guidance on when to use what tools. Designers should be aware of what kind of data will be tracked and possibly suggest the most appropriate way of analyzing it, especially for the time-series data.

Smartphone technology is focusing on accumulating data from different apps which the users uses and employing machine learning techniques to personalize the experience of interacting with the device. This direction could be adopted by the designers of tracking applications to personalize the design methodology based on the individual's lifestyle. For example, if a person chooses to track their sleep quality based on their physical activity, the tracking application may access their calendar program, to decide which design methodology is appropriate based on their schedule. Another example is to use the geo-location of the user to find an outlier data point when the person is travelling, thus accounting for changes in time zone and addressing possible effects on sleep such as jet lag.

Further, 13 students tracked similar activities to one another such as sleep quality. Designers should create special utility tools for commonly tracked activities to ease the users' experience and to get started with tracking. In addition, designers should focus on both basic and advanced means of analyzing the data, which would be subjected to the individual's interest and the variables which they want to track. It can also be seen from the study that ten experiments had visualizations as their basic way of analyzing the data and drawing a conclusion. Hence, if the analysis of an experiment is coupled with visualizations, then the results will have a huge impact on the decision of whether to go through a purposeful behavior change or not.

CONCLUSION

In this paper, we described a meta-analysis of twenty N=1 experiments where students in a Computer Science Human-Computer Interaction seminar were introduced to experimental design and given a structured assignment to perform a personal informatics experiment. They performed the experiments with a scientific mindset, using the framework of hypotheses, independent and dependent variables, and statistical testing. This procedure allowed us to observe what happens when people go beyond a typical self-observation paradigm into an experimental one, which is itself a difficult scenario to reproduce outside the classroom due to the lack of consistency and background of the self-trackers. The students typically applied an interrupted time-series experimental design to manipulate different variables in their experiments.

We compare the students' experimental designs, methods, and outcomes to those of self-motivated Quantified-Selfers. Based on this comparison, we propose that one month is not long enough to reach conclusive results, that iterations on the data collection and analysis stage can be beneficial to experimenters, and that an exploratory stage at the beginning of the experiment could lead to a better design of the study later on. Based on our findings, we propose an iterative self-experiment design, which is comprised of an exploratory stage with a naturalistic design, followed by iterations on the actual data collection and analysis stage. We believe that this design model could be helpful to both future personal informatics experimenters and designers of tools.

Our work contributes to the broader understanding of personal informatics, where prior work has emphasized the importance of self-experimentation, but admitted that self-trackers often lack the background to run a rigorous scientific-like experiment. We learn what happens when people are given the basic understanding of experimental design and guidance to run a personal informatics experiment, and how this may affect the way we teach people to better understand themselves. This is one of the ways in which we can personalize the study of people from broad population-level studies which are often cast too widely to be useful to individuals, to N=1 studies which are immediately relevant and targeted to oneself.

REFERENCES

1. 2014. Brain Basics: Understanding Sleep. (2014). Retrieved September 24, 2015 from `http://ninds.nih.gov/disorders/brain_basics/understanding_sleep.htm`.
2. 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 (2013), 30:1–30:27. `DOI:http://dx.doi.org/10.1145/2503823`
3. Jesse A. Berlin. 2010. N-of-1 clinical trials should be incorporated into clinical practice. *Journal of Clinical Epidemiology* 63, 12 (2010), 1283–1284. `DOI: http://dx.doi.org/10.1016/j.jclinepi.2010.05.006`
4. 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`
5. Nelson Cowan. 2008. What are the differences between long-term, short-term, and working memory? *Progress in brain research* 169 (2008), 323–338.
6. Joan Didion. 1968. On Keeping a Notebook. (1968). Retrieved September 24, 2015 from `https://penusa.org/sites/default/files/didion.pdf`.
7. 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`
8. Garabed Eknoyan. 1999. Santorio Sanctorius (1561–1636)–founding father of metabolic balance studies. *American journal of nephrology* 19, 2 (1999), 226–233.
9. 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`
10. Joshua Foer and Michel Siffre. 2008. Caveman: An Interview with Michel Siffre. *Cabinet* 30 (2008).
11. BJ Fogg. 2015. Tiny Habits. (2015). Retrieved September 24, 2015 from `http://tinyhabits.com/`.
12. 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.
13. Eric B Hekler, Winslow Burleson, and Jisoo Lee. 2013. A DIY Self-Experimentation Toolkit for Behavior Change. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.
14. Jim A Horne and Olov Ostberg. 1975. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International journal of chronobiology* 4, 2 (1975), 97–110.
15. 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.
16. Quanfied Self Labs. 2012. About the Quantified Self. (2012). `http://quantifiedself.com/about/`
17. 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`
18. Renae Low. 2012. Modality Effect on Learning. In *Encyclopedia of the Sciences of Learning*, NorbertM. Seel (Ed.). Springer US, 2295–2298. `DOI:http://dx.doi.org/10.1007/978-1-4419-1428-6_256`
19. Barry Marshall and Paul C Adams. 2008. Helicobacter pylori: A Nobel pursuit? (2008). Retrieved September 24, 2015 from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2661189/`.
20. 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.

21. Ernesto Ramirez. 2015. QS Access: Ian Eslick on Personal Experimentation. (2015). Retrieved September 24, 2015 from `http://quantifiedself.com/2015/01/qs-access-ian-eslick-personal-experimentation/`.

22. 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.

23. Seth Roberts. 2010. The unreasonable effectiveness of my self-experimentation. *Medical hypotheses* 75, 6 (2010), 482–489.

24. 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`

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