QuantifyMe: An Automated Single-Case Experimental Design Platform

Akane Sano(✉), Sara Taylor, Craig Ferguson, Akshay Mohan, and Rosalind W. Picard

Massachusetts Institute of Technology, Media Lab, Cambridge 02139, USA
{akanes,sataylor}@media.mit.edu

Abstract. We designed, developed, and evaluated a novel system, QuantifyMe, for novice self-experimenters to conduct proper-methodology single-case self-experiments in an automated and scientific manner using their smartphones. In this work we evaluate its use with four different kinds of personalized investigations, examining how variables such as sleep duration and regularity, activity, and leisure time affect personal happiness, stress, productivity, and sleep efficiency. We describe lessons learned developing the system, and recommendations for its improvement, as well as its potential for enabling personalized insights to be scientifically evaluated in many individuals.

Keywords: Single-case experimental design · Mobile health
Self-experiment · Self-tracking

1 Introduction

Mobile devices today have become nearly ubiquitous full-time extensions of an individual, enabling unprecedented data collection through a combination of mobile and wearable sensors. The collected data can provide an individual with a unique opportunity to determine behavioral patterns and habits about themselves [7]. A common way for individuals to use these data is to compare the logged values of a variable of interest with its recommended values (e.g., sleep, physical activity, and calorie intake), and then adjust daily behavior accordingly. However, these recommendations made by agencies like Center for Disease Control, are averages and ideal values may vary significantly in individuals. Although average values are useful, they do not provide the personalized insights that are best suited for an individual to make optimal behavioral choices.

Our system, QuantifyMe, creates a framework that allows people to find their personal optimal behavioral variables (e.g., bed time or physical activity) to achieve their goals (e.g., productivity or happiness) based on evidence-based experimentation. This is done through a single-case experiment design methodology — a methodology that allows for comparison within an individual instead of between groups. In achieving these aims we are closer to including the general public in dramatically scaling and personalizing the study of daily behaviors.

A. Sano and S. Taylor – Contributed equally to this work.

P. Perego et al. (Eds.): MobiHealth 2017, LNICST 247, pp. 199–206, 2018.
https://doi.org/10.1007/978-3-319-98551-0_23

2 Related Work

2.1 Quantified Self

Mobile applications and wearable sensors have enabled individuals to collect personal real-time behavioral and physiological data such as physical activity, sleep [10], and diet [11] and use these data to infer their daily activity patterns. This also allows for self-experimentation to make better choices related to their lifestyle, health, and productivity goals. However, using these data to understand optimal values of variables and causal relationships in behaviors like sleep, physical activity, and caloric intake often lacks scientific rigor [3]. Recent mobile applications have provided great tools supporting more systematic and personalized self-experimentation [5,8,13,14]. However, many of these tools do not provide a structured methodology to aid non-scientist quantified selfers, and/or they have not been validated on usability and effectiveness.

2.2 Single-Case Experimental Design

Randomized Control Trials (RCTs) are considered a gold standard in determining whether a causal relationship exists between a specific intervention and observed outcome [16]. However, traditional RCTs operate only across groups, and are unable to provide individual insights [1]. First proposed by Neuringer [17], single-case experimental design provides a methodology that allows researchers to evaluate the effectiveness of an intervention on an individual. An individual serves as his or her own control and is subjected to different experimental conditions at different time periods [1,15]. This contrasts with a group-based design in which outcomes are compared between groups, with each group receiving a specific experimental condition.

3 Survey Study for Understanding Users' Interest in Self-experimentation

Before developing our smartphone app and system, we conducted an online survey study to gauge users' interest in self experimentation and to find which of 32 self-experiments they are interested in. The possible output variables of these proposed experiments (happiness, stress, productivity, sleep efficiency) were chosen from common indicators of wellbeing. The possible independent variables (active time, steps, sleep duration, bed time, meditation duration, outdoor time, fun time, attending a religious service) were chosen based on the ability to be measured with the Jawbone wearable sensor or because they are relatively easily controllable and actionable [14].

A total of 233 individuals completed the survey (90%: 18–24 years old, 5%: 25–29 years old, 5%: over 30 years old). Based on the results of the survey, we decided to focus on four self-experiments — one for each outcome variable: (1) How does my leisure time affect my happiness? (2) How does my activity level affect my sleep efficiency? (3) How does my nightly sleep affect my productivity? (4) How do inconsistent bedtimes affect my stress level? While we selected these four experiments for their popularity for use in the first version of the app, the QuantifyMe system was designed to be flexible to accommodate many different kinds of self-experiments, not just these four.

4 QuantifyMe System for Self-experimentation

The QuantifyMe system consists of three parts: a backend Django application, an Android App, and a Jawbone UpMove fitness tracker. The system could be expanded to other fitness trackers and smartphone platforms.

4.1 Single-Case Experimental Design on QuantifyMe

A traditional suggested design for single-case experiments is an $ABAB$ design, where the A phase corresponds to the baseline, and the B phase corresponds to the intervention period. This design can be modified as a non-terminated sequential intervention $AB_1B_2B_3$ design to see the relationship between different magnitudes of the intervention and their outcomes [18]. This is best suited to our system as we are looking to determine the optimum magnitude of the independent variable. Therefore, we implemented a four-stage design (1 baseline stage and 3 intervention stages) in order to help users determine optimal behaviors with each stage including 4–7 days of data points as suggested by [1].

We quantized behaviors into five zones for each experiment (see Table 1). These target behaviors were predetermined by examining common behaviors based on another study [19].

The "randomized" ordering of target goals was chosen as follows: Stage 1 was a baseline measure. Because a choice needed to be made, we settled on having the middle stage (O2) be the last stage. We also decided to include at least one increase in the target behavior and one decrease in the behavior.

As an example of intervention order, if a user's average sleep duration during Stage 1 (baseline period) is 6.75 h (i.e., within O1), the user would be instructed to sleep 8.5 hr, 6.5 hr, and 7.5 hr during stages 2, 3, and 4, respectively. However, if the mean of the user's sleep duration during Stage 1 was 8.75 h

Table 1. Definitions of target zones of behaviors for each experiment

Zone	Number of steps	Bed time variability	Sleep duration	Leisure time
Under	< 6,500 steps	< ±15 min	< 6 hr	< 15 min
O1	8,000 (6,500–9,500)	±30 min (±15–±45)	6.5 hr (6–7)	30 min (15–45)
O2	11,000 (9,500–12,500)	±60 min (±45–±75)	7.5 hr (7–8)	60 min (45–75)
O3	14,000 (12,500–15,500)	±90 min (±75–±105)	8.5 hr (8–9)	90 min (75–105)
Over	> 15,500 steps	> ±105 min	> 9 hr	> 105 min

(i.e., within O3), the user would be instructed to sleep 6.5 h, 8.5 h, and 7.5 h during stages 2, 3, and 4, respectively. The methodology of imposing sleep targets adds more structure and validity to determining a causal relationship than does simply correlating how long a user chooses to sleep each night with how they feel the next day.

4.2 QuantifyMe Android App

The Android app was designed with the goal of letting the user easily enter data for daily check-ins, while also allowing the user to check on the status of their current experiment. When the user first opens the app after installation, it prompts them for demographic data and asks them to select an experiment. The app connects their Jawbone account to our system's account.

Every morning during the experiment, the user is reminded to check-in on the app and fill out a short daily survey. This survey asks about the amount of leisure time in the past 24 h, along with happiness, stress and productivity levels using 7-point likert scales (not at all — extremely). Finally, the app reminds the user to sync their Jawbone wearable sensor to Jawbone's Up App (syncing takes a few seconds).

After the user has checked-in for the day, the app presents the user with a screen that lets them view their daily goal and experiment progress during that stage (see Figs. 1b and c). In particular, the user is able to see her recorded behavior for all of the days she has been in that stage.

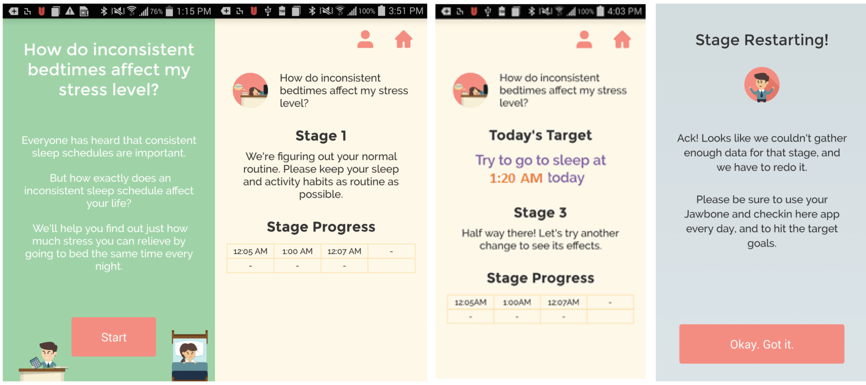
If a user has failed a stage in the experiment, they are shown a message prompting them to restart the stage (see Fig. 1d). Once an experiment has been completed successfully, the user is shown a success screen with their end results, and the experiment's results are added to their history, which they can view from the daily goals screen at any time.

5 App Evaluation Study

After the protocol was approved by the Institute Review Board, we conducted a 6-week pilot study to evaluate the new QuantifyMe application with 13 participants (4 male, 9 female age: 18–27). All participants filled out a Big Five Personality Inventory [12] (not analyzed in this paper) and then chose a self-experiment they liked, which was continued for the 6 weeks. At the end of the study, the participants filled out a post-study survey including a System Usability Scale [2] and questions about their favorite/least favorite parts of the app.

5.1 Results

Self-Experiment Selection. Among the four experiments on the app, 5 people chose "effect of sleep duration on productivity," 4 people chose "effect of leisure on happiness," 2 people selected "effect of sleep variability on stress," and 2 selected "effect of steps on sleep efficiency". This distribution matches that of the survey we conducted before designing QuantifyMe (see Sect. 3).



(a) (b) (c) (d)

Fig. 1. Screenshots of the QuantifyMe Android App: (a) information provided before starting an experiment, (b) example of instructions during stage 1 (baseline period), (c) example of instructions during stage 3, and (d) example of stage restart notification

Adherence. During the 6-week study, three participants dropped out for various reasons including phone malfunction, side-effects of an un-related medication, and the self-experiment (inconsistent bedtimes and stress) being too difficult to complete. Thus, 10 participants completed the study (i.e., used the QuantifyMe app for 6 weeks); however, only one participant successfully completed a full four-stage scientific self-experiment during the 6-week study.

The average adherence rate for checking-in (i.e., *check-in adherence*) to the QuantifyMe app was 75.8%. *check-in adherence* remained stable throughout the study, and decreased rapidly after the study ended. *Check-in adherence* rates varied widely with four participants checking-in on less than 65% of the days and 3 participants checking-in on more than 90% of the study days. In comparison, we found that on average participants adapted their behaviors to be in the target range on 22.5% of the study days (i.e., *objective adherence*). This lack of adherence to self-experiment instructions was the main reason why only one participant completed a self-experiment in the 6-week period. In other words, many participants had trouble adjusting their behaviors to match the self-experiment instructions (Fig. 2). This resulted in many stage restarts because we required participants to check-in at least 5 out of 7 days and be within the target behavior range for at least 4 of the 7 days.

Post-study Survey. The System Usability Scale (SUS) results presented are from 10 study participants and excludes three who had technical difficulties with use of the app on their phone. The SUS showed a mean of 71 (out of 100) with a standard deviation of 17 (Fig. 3). A SUS score of 68 is considered average [20]. Therefore, our system scored slightly above the average system for usability.

Post-study comments were analyzed from all 13 participants. Seven out of 13 study participants indicated that the daily survey and stage progress allowed

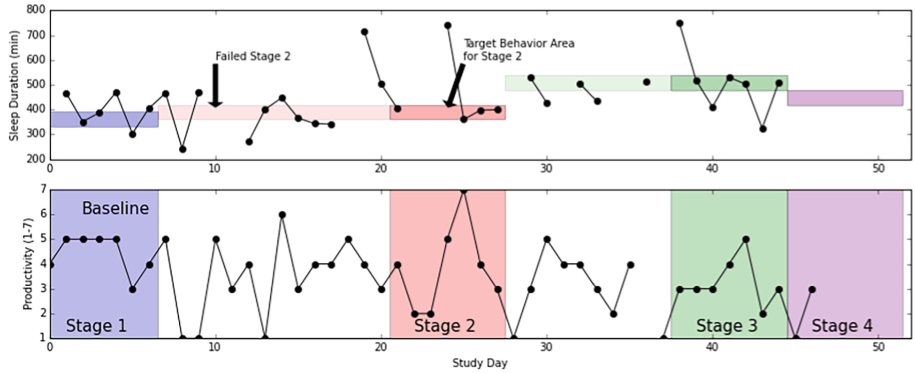


Fig. 2. Example of a participant's self-experiment. The top panel shows their behavior manipulating the independent variable: sleep duration. The bottom panel shows the outcome variable: self-reported productivity. The shaded areas (blue, red, green, and purple) mark each of the 4 stages of the study. In the top panel, we can see that these shaded areas also display the bounds of target behavior. The lighter shaded areas serve as a reference for the target behaviors when the experiment stages had to be restarted.

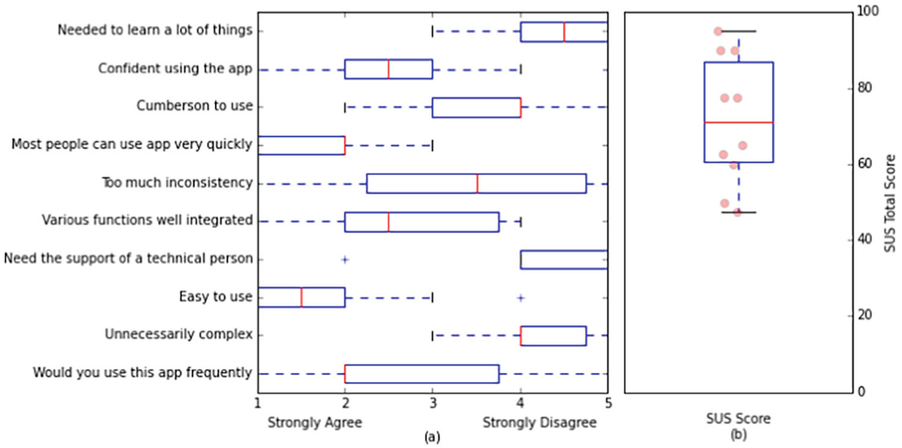


Fig. 3. Results of the system usability scale. Figure (a) shows the distribution of responses for each question and (b) shows the distribution of the SUS total score.

them to be more aware of their behavioral patterns and prompted introspection of their lifestyle. Most study participants indicated their disappointment in not being able to complete the full-four-stage self-experiment during the 6-week study. We also asked about the least favorite aspects of the app. One participant said, "It was hard to follow the instruction about sleep schedule because it was more than I am used to and I wasn't able to plan my schedule around that". Another participant said, "Worrying about when I wasn't able to complete the app instructions" was their least favorite aspect. These comments were consistent with the behavioral lack of success in changing their independent variable.

6 Discussion

This work has several limitations. We assumed that self-experimenters would have enough motivation to maintain the different behaviors being tested. Clearly, better mechanisms to encourage behavioral compliance are needed. Participants suggested it would be valuable to have additional information and motivation about the single-case design methodology such as the number of stages and notifications of why a particular experimental stage would need to be restarted if they didn't reach the target for the independent variable. In the future we recommend that providing study participants with the upcoming targets multiple days ahead and clarifying what flexibility is and is not going to be a set back in the study may allow them to plan for the interventions in advance and increase adherence.

Our population was a set of busy healthy individuals who think they may be interested in trying to find optimal behaviors; they are not likely to be as motivated as an unhealthy population who seeks treatment. We also designed multiple targets they had to meet (instead of sticking with one intervention), adding additional challenges to engaging our healthy population.

Because most participants in our studies were university students who are savvy about technology, experimental design and statistics, these results might not generalize to other populations. Also the participants may have been fairly homogeneous in terms of experiments they may be interested in. One sign supporting this was that the most popular experiment related to the amount of sleep needed. However, all of the participants in the pilot study were novice self-experimenters, which suggests that other tech-savvy novice self-experimenters may encounter similar challenges.

7 Conclusions

In this work, we designed, developed, and evaluated a novel system for users to conduct single case self-experiments in a scientific and automated manner. The QuantifyMe system was designed to create a framework for novice self-experimenters to find their personal optimal behavioral variables to achieve their goals by automating the single-case experimental design process within a mobile application. In a pilot study, we found that although target-behavior compliance was low, our participants still expressed interest in having such an system to determine how to personalize optimal behaviors. Future versions of the QuantifyMe system should include methods of increasing compliance via maintaining motivation on a daily basis and better preparing participants to be able to hit the target behaviors.

References

1. Barlow, D.H., Hersen, M.: Single Case Experimental Designs. Pergamon Press, New York (1984)
2. Brooke, J.: SUS - A quick and dirty usability scale. Usability Eval. Ind. **189**, 4–7 (1996)
3. Choe, E.K., et al.: Understanding quantified-selfers' practices in collecting and exploring personal data. In: CHI 2014, pp. 1143–1152 (2014)
4. Cordeiro, F., et al.: Rethinking the mobile food journal. In: CHI 2015, pp. 3207–3216 (2015)
5. Daskalova, N., et al.: SleepCoacher: a personalized automated self-experimentation system for sleep recommendations. In: Proceedings of 29th Annual Symposium User Interface Software Technology, pp. 347–358 (2016)
6. Dugard, P.: Randomization tests: a new gold standard? J. Context Behav. Sci. **3**, 65–68 (2014)
7. Epstein, D.: Personal informatics in everyday life. In: UbiComp 2015, pp. 429–434 (2015)
8. Evans, B.: PACO Personal Analytics Companion (2017). https://quantifiedself.appspot.com/
9. Halko, S., Kientz, J.A.: Personality and persuasive technology: an exploratory study on health-promoting mobile applications. In: Ploug, T., Hasle, P., Oinas-Kukkonen, H. (eds.) PERSUASIVE 2010. LNCS, vol. 6137, pp. 150–161. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13226-1_16
10. Jawbone Inc., Jawbone Up App (2017). https://jawbone.com/
11. MyFitnessPal Inc., MyFitnessPal Inc. (2017). https://www.mytnesspal.com/
12. John, O.P., Srivastava, S.: The Big Five trait taxonomy: history, measurement, and theoretical perspectives. Handb. Personal Theory Res. **2**, 102–138 (1999)
13. Karkar, R., et al.: TummyTrials: a feasibility study of using self-experimentation to detect individualized food triggers. In: CHI 2017 (2017)
14. Karkar, R.: A framework for self-experimentation in personalized health. J. Am. Med. Informatics Assoc. **23**, 440–448 (2016)
15. Lillie, E.O.: The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per. Med. **8**, 161–173 (2011)
16. Meldrum, M.L.: A brief history of hte randomized controlled trial: from oranges and lemons to gold standard. Hematol. Oncol. Clin. North. Am. **14**, 745–760 (2000)
17. Neuringer, A.: Self-experimentation: a call for change. Behaviorism **9**(1), 79–94 (1981)
18. Newman, F.L., Wong, S.E.: Progress and outcomes assessment of individual patient data: selecting single-subject design and statistical procedures. In: Use of Psychological Tests for Assessing Treatment Outcomes, vol. 1 General considerations, pp. 273–289, 3rd edn. (2004)
19. Sano, A.: Measuring college students' sleep, stress, mental health and wellbeing with wearable sensors and mobile phones. Ph.D. Dissertation. Massachusetts Institute of Technology (2016)
20. Sauro, J.: SUStisfied? Little-known system usability scale facts. User Exp. 14–15 (2011)