Comparing Qualitative Evaluation Methods for Improving User Engagement with Mobile Health Applications

Katherine S. Finnerty[1], Carolina Ruiz, Ph.D.[2], Bengisu Tulu, Ph.D.[2], Qiaoyu Liao, BS[2], Jessica Oleski, MA[3], Sherry Pagoto, Ph.D.[3]

[1]Amherst College, Amherst, Massachusetts; [2]Worcester Polytechnic Institute, Worcester, Massachusetts; [3]University of Massachusetts Medical School, Worcester, Massachusetts

Abstract

As obesity has become a more widespread issue both in the United States and globally, weight loss has become a major focus of mobile health. SlipBuddy is an Android application that collects data from its users to learn patterns from and make predictions about their overeating behavior. A constraint on the development of applications such as SlipBuddy is that evaluating their quality often relies on large, resource-intensive studies with many users over a long period of time. This project investigates how to best address this constraint by leveraging limited user study data with data analysis techniques such as text mining and association rule mining. The result is that using an objective analysis based on captology frameworks in conjunction with semi-structured feedback from users can result in concrete, actionable recommendations for improvement in mobile health application development.

Introduction

“SlipBuddy” is an Android application that uses machine learning to track and predict overeating episodes[1]. The application defines “overeating” as any eating that the user feels resulted in consuming (food or drink) more than planned at a meal or between meals. Users log three status updates per day (morning, afternoon, and evening check-ins) along with any overeating episodes, or “slips” throughout the day. The application predicts when the user is at risk of overeating based on their past behavior and to offer behavioral interventions to prevent overeating. The morning status update collects the user’s stress and hunger levels, number of hours of sleep, how well-rested the user feels, and the user’s weight for that morning. The afternoon status update asks about stress level. The evening status update also asks about stress level and reminds the user to report any overeating episodes that they have not yet reported.

SlipBuddy has undergone various development stages, including a pilot study conducted with 16 participants who used the application for two months. The pilot study showed high user participation during the first month (90.27%) but declining user participation in the following month (63.3%). There was also evidence to suggest incomplete reporting of overeating episodes. In this paper, we present our approach to objectively measure how well SlipBuddy engages its users using the small pilot study sample to answer our research questions: How well does SlipBuddy persuade its users to continue using the app, and what are the app features that require improvement?

Captology is defined as the study of interactive computer technology designed to change user attitudes or behaviors[2]. Three frameworks commonly used to design and evaluate such persuasive technologies are Persuasive Systems Design Model (PSDM), the Elaboration Likelihood Model (ELM), and the Functional Triad[2]. The PSDM outlines a framework within which developers can work to develop an interactive persuasive technology[3]. The ELM describes two routes within which users can process information: the central route, which is more cognizant, and the peripheral route, which relies on subconscious cues[4,5]. The Functional Triad distinguishes three ways in which users view their relationship with technology (computers as tools, as medium, and as social actors), and how developers can take advantage of these relationships[2].

In this paper, we present (1) our approach to evaluation of SlipBuddy’s persuasive performance based on these frameworks where we determine areas in which the app performed well and areas in which the app could improve, (2) how we conducted text mining analysis of interview transcripts from the pilot study to support our findings from theoretical evaluation process, (3) how we investigated the level of awareness users have about their behavioral patterns by comparing analysis of self-reflection during the interviews to analysis of association rules mined from the self-reported data collected by the app. We argue that theoretically and empirically grounded approach we utilized helps determine specific app enhancements that can improve its efficacy.

Methods

Data Collection: The SlipBuddy pilot study began in November 2015 and consisted of two month-long phases, the Assessment Phase (n=16) and the Feedback Phase (n=14)[1]. Participants were over the age of 18, owned an Android smartphone running version 4.0 or later, were either overweight or obese according to their BMI, had familiarity with mobile applications, and were interested in losing weight. Those who did not use their phone daily, could not consent, or who were pregnant or prisoners were excluded. Recruitment occurred first from the local community and then from the entire country to maximize generalizability.

In the Assessment Phase (Phase 1), participants were asked (1) to complete status updates three times a day and (2) to report any overeating episodes and to provide information about the circumstances of those episodes. At the end of this phase, interviews were conducted to discuss the features of the application, their use and suggestions for improvements. The Feedback Phase (Phase 2) started after a two-month break where developers used predictive analytics to analyze Assessment Phase data and to implement interventions. During Phase 2, participants used the application in the same way as before. They also received interventions when their behavior indicated that they were at a higher risk of overeating. At the end of Phase 2, participants took part in another interview about the application and adherence to its use. Some participants did not receive interventions because the combination of triggering factors identified on their Phase 1 data were not observed during Phase 2, so they could not provide their opinions on this new feature.

Data Analysis: *First,* we analyzed SlipBuddy based on each persuasive technology framework independently. Each framework consists of two or three main sections, each containing multiple factors. The lead author evaluated SlipBuddy based on each relevant factor subjectively. The evaluation results were then discussed with the research team and finalized in meetings. Second, we analyzed the pilot study interviews (30 total interviews with 16 participants) using a Bag of Words text mining approach[6]. We excluded from the analysis "stopwords" (e.g., "a", "the", "that") as well as other words expected to appear frequently in the interview transcripts, including: "slip," "eat," "think," and "feel." Next, we aggregated participant responses from each phase to find frequent unigrams, bigrams, and trigrams (one, two, and three word sets, respectively). We compared the results of each phase's analysis and constructed commonality and comparison plots for Phase 1 and Phase 2 aggregate responses. Later, we grouped responses to each of the interviewer's questions." We conducted the same term frequency analysis on these files as we did on the aggregate phase data. We reviewed the results to identify trends and to determine areas of success and areas for improvement**.** Third, we used association rules to look at data-based patterns in individual users' overeating. An association rule is a collection of data items that are likely to co-occur in the data[7]. We used the Apriori algorithm to obtain association rules with minimum support of 40%, collections of factors that appeared together in at least 40% of the days when the user reported overeating[8]. Factors under consideration included the user's activity, location, time, stress level and hunger levels at the time of the overeating episode, as well as all data corresponding to that day's status updates. We reviewed participants' responses to questions about their overeating behavior to extract a similar list of factors and check for any similarities or dissimilarities between these factors and factors in the association rules, showing insight on users' awareness of their own behavior. Triangulating the results from each analysis revealed themes supported across both the subjective and objective data.

Results

The Persuasive Systems Design Model (PSDM) framework indicated that SlipBuddy has sufficient functionality and operates well in the persuasive context of overeating reduction (Tables 1-2)[3]. The application's simplification of weight loss tracking works well within the model. However, the application lacks support in the Dialogue Support and Social Support feature categories (Table 3).

Table 1. The first stage of the PSDM: understanding key issues of persuasive systems[3].

Persuasive Systems Topic	**SlipBuddy Context**
Information technology is never neutral.	SlipBuddy is explicit about its weight loss motivation.
People like their worldviews to be consistent.	SlipBuddy works with users' goals and ideas.
Direct and indirect routes are key persuasion strategies.	SlipBuddy uses indirect routes throughout the interface and direct routes in the intervention.
Persuasion is often incremental.	SlipBuddy encourages change over time.
Persuasion should always be open.	SlipBuddy does not deceive users about its function.
Persuasive systems should aim at unobtrusiveness.	SlipBuddy interventions only appear after check-ins, and notifications are not overly obtrusive.
Persuasive systems should be useful and easy to use.	SlipBuddy has a simple user interface and functionality.

Table 2. The second stage of the PSDM: analysis of the persuasion context[3].

Element of the Persuasion Context	SlipBuddy Context
Intent	Persuasive intent comes from all involved parties: the users and creators both want to change users' behavior to reduce overeating.
Event	Overeating context addressed through factors such as stress and sleep levels.
Strategy	Machine learning takes advantage of technology-dependent features.

Table 3. The third stage of the PSDM: evaluation and development of system qualities[3].

Category of System Qualities	SlipBuddy Context
Primary Task Support	Most of SlipBuddy's features focus on simplifying the task of tracking and preventing overeating and helping the user through the process.
Dialogue Support	The only dialogue is in interventions.
System Credibility Support	The best way to measure perceived credibility is to track performance over an extended period of time[2], so more work and time are needed.
Social Support	Currently, there is no social support except for the intervention dialogue.

The Elaboration Likelihood Model (ELM) examines whether users will take a central or peripheral route in cognitive processing (Table 4)[4,5]. The central route, which has more long-term effects, relies on high intrinsic motivation and ability to process information, while the peripheral route relies on implicit cues and associations. When the elaboration likelihood is high, the central route occurs. Otherwise, the peripheral route occurs[4]. Personalization is an effective tactic to appeal to users in the peripheral route[5]. Even the mere impression of tailoring can increase persuasiveness. Currently, the main instance of personalization in SlipBuddy is the intervention. The phrase "according to your data" makes it clear that the intervention is specifically for the user.

Table 4. The persuasive process according to the ELM[4]. A high intrinsic motivation and ability to process information makes users more likely to take a central route in cognitive processing.

Persuasive Process Step	SlipBuddy Context
Motivation to Process Information	SlipBuddy users engage with the application wanting to lose weight, so they will be receptive to process information from the weight loss application.
Ability to Process Information	The application is minimal and simple enough that it should be accessible to all users.
Cognitive Processing	Based on the above rows, SlipBuddy users are more likely to take the central route, but features should be present for both central and peripheral routes.

Evaluating SlipBuddy within the Functional Triad[2] revealed that SlipBuddy works better as a tool than as a social actor. SlipBuddy's Primary Task Support features in the PSDM also enhance SlipBuddy's ability to act as a useful tool. The Kairos factor, which is defined as the most opportune moment to intervene to a user, accounts for the time of day along with the user's location, routine, goals, task, and any other relevant factors[2]. SlipBuddy does not take full advantage of the Kairos factor. The application only provides interventions immediately after check-ins. SlipBuddy addresses most of the factors of "Computers as Tools," but there is room for improvement for the factors of "Computers as Social Actors" (Table 5).

Table 5. SlipBuddy's use of social persuasive strategies according to the Functional Triad[2].

Social Persuasive Strategy	SlipBuddy Context
Physical Cues	SlipBuddy has no physical entity or character.
Psychological Cues	SlipBuddy has no displayed emotion or personality.
Language	SlipBuddy uses lingual support in interventions.
Social Dynamics	SlipBuddy invokes some reciprocity by processing the users' data to help them.
Social Roles	SlipBuddy has potential to adopt the role of a friend.

The text mining analysis using a Bag of Words approach showed that participants agreed the application was easy to use, with “easy use” being by far the most frequent bigram in aggregate responses to “What did you like about the application?” (Figure 1). Analysis of other resulting word clouds (Figure 1) revealed that participants did not like the discomfort from being aware of their overeating and that participants relied on calorie counts to track overeating. Many participants expressed a lack of clarity in what certain terms within the application meant, such as quantifying stress or defining overeating.

Figure 1. Text Mining Results - On the left, a bar chart showing most frequent bigrams among participant responses to the question “What did you like about the application?” after Phase 2 of the pilot study. In center, a word association cloud from Phase 1 aggregate transcripts based on “dislike.” On the right, a word frequency cloud from Phase 2 responses to “How did you define overeating?”

A common theme in the participant feedback across phases and questions was the desire for more notifications and more interaction in general. This offers evidence that users would be receptive to a more involved, talkative application as described based on the captology analysis. Participants wanted more specific suggestions for healthy behavior, such as eating healthy snacks or going to sleep early. This is an opportunity to implement more dialogue for the application. Some insights from the text mining analysis were context-specific and not concretely detected in the captology framework analysis. For example, participants mentioned that the “View History” feature for overeating episodes was not very useful. Participants reported that they were not weighing themselves every day as the application asked, an issue that would be impossible to find with the captology analysis. Participants also mentioned the inconvenient timing of the check-ins. This relates to a recommendation from the captology analysis: user customization. Allowing users to adjust their check-in times would address the last two issues raised by participants.

For text mining, we focused on mining participants’ vocalized opinions. Meanwhile, participants’ vocalized descriptions of their overeating behaviors help create a comparison between their perception of their overeating and the patterns detected by the analysis of their actual overeating data in the app using Apriori algorithm (Table 6). When extracting factors from participants’ responses, we prioritized both matching the language of association rules and accurately reflecting meaning. This process was completed for each of the 16 individuals. However, due to some participants having too few data entries and some having too many association rules with high support levels such that the rules were meaninglessly repetitive, we could only obtain results for 10 participants.

For most participants, the association rules revealed similar trends to our findings from earlier analysis. For example, Participant 11 states that they tend to overeat while snacking at home during the nighttime after work. This matches up with three of their association rule factors: episode time: evening, snacking, and at home. However, we saw contradictory results for Participants 1 and 5. They state that high stress makes them more likely to overeat, whereas the association rule data shows the opposite: they are more likely to overeat under low stress. There are two possible causes for this conflict. The first is that the users incorrectly verbalize their behavior, perhaps relying on the belief that most people eat when stressed. The second is that low numbers on the stress scale still correspond to high perceived stress to the user.

Table 6. Factors associated with overeating extracted from association rule mining of overeating data and user interviews in which users described their overeating. Only factors in multiple association rules are shown.

Participant ID	Association Rule Factors	Participant Interview Factors
1	working/studying, eating breakfast, moderately full, low morning stress, medium stress, low PM stress, weight gain, well-rested	high stress, snacking, during nighttime
2	moderately full, medium stress, somewhat well-rested, usual amount of sleep	high stress
4	moderately full, medium stress, usual amount of sleep, at home, somewhat rested, episode time: evening	snacking, socializing, not hungry
5	low stress, moderately full, at home, well-rested	high stress
7	moderately full, usual amount of sleep, low stress, no weight change	snacking, socializing
8	snacking, low stress, usual amount of sleep, moderately full in AM	unplanned eating
10	low PM stress, high AM stress, very hungry, at home	unhealthy foods, during meals, socializing
11	episode time: evening, snacking, low stress, at home, moderately full, no weight change	nighttime, snacking, at home
14	medium stress, well-rested, moderately full in the morning, well rested, high PM stress, usual amount of sleep, episode time: late afternoon	socializing, less sleep than usual
16	medium stress, moderately full in the morning, somewhat well-rested, no weight change, usual amount of sleep	high stress, nighttime

Discussion

The captology and text mining analysis showed that SlipBuddy's main strengths lie in its simplicity and personalized interventions. Although more improvements and customization are possible, the functionality already holds persuasive power through its navigable interface and credible analysis. However, there remain three key areas that developers could improve upon to help retain user engagement.

The first area for improvement based on this analysis is intervention timing. Implementing interventions outside of check-in times will increase functionality. The second area is tailoring. SlipBuddy already prominently uses personalization in its machine learning-based interventions, and making this personalization clearer to users will increase its persuasive power. Also, implementing more user settings will help make the user have more control over their experience (Figure 2). This sense of ownership will encourage them to return to the application. The final, most important area is SlipBuddy's capacity to play a social role. By increasing dialogue, SlipBuddy can motivate and encourage users. In addition to functionality, the application can provide a positive social experience (Figure 2).

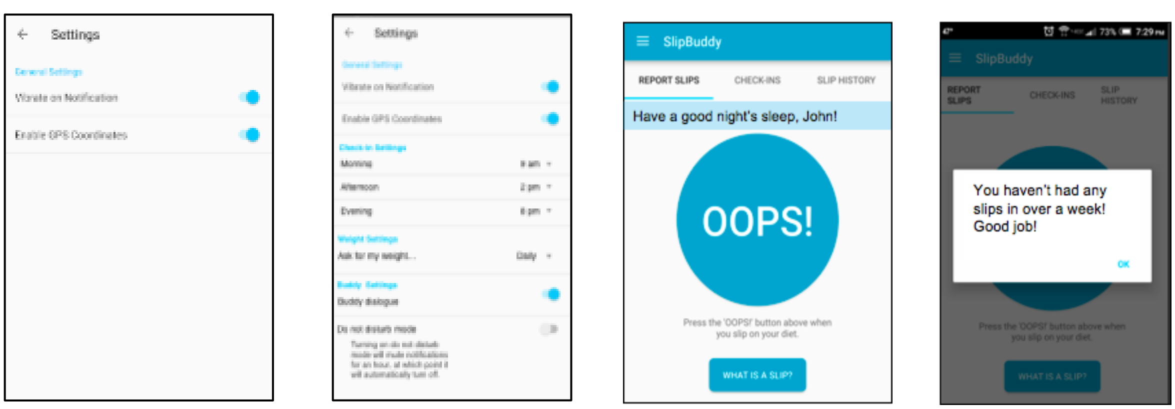

Figure 2. On the left, the current settings page. On the center left, a potential new settings page. On the center right, a mock-up of a header that might appear during the evening to encourage a user to get more sleep. On the right, a mock-up of a notification that might appear after a user's status update, taking better advantage of the Kairos factor.

There are various categories of features which only one of the methods could address. The captology analysis's generality means that it cannot identify problems pertaining to SlipBuddy in particular. However, the captology analysis emphasizes nuances in user psychology that participants are unlikely to identify when describing their experiences using the application in an interview. Since the captology analysis and the text mining analysis both have their limitations, they work best when used in conjunction with each other.

The comparison of overeating factors extracted by association rule mining from the application's collected data and those extracted directly from the interviews was useful in pointing out that participants' perceptions of their behavior agree in many cases with their self-reported data but not in all cases. In addition to recommendations from the captology analysis and the text mining, the association rule mining yields another recommendation. In the column of association rule factors in Table 6, every participant except for Participant 10 has some association rule factor that is moderate or medium, indicating values that appear in the middle of the input scale for users. These factors are not only less informative than factors corresponding to the more extreme ends of the input scale; they may also be less accurate. When the user completes a status update, SlipBuddy sets the answers to several default values, all in the middle of the input scale. Thus, if users feel lazy, they may ignore a question and leave it at the default value. In these circumstances, the central input values are not informative in terms of predicting user overeating. Changing this system will help increase SlipBuddy's effectiveness in developing overeating interventions.

Conclusion

Exhaustive user studies require extensive amounts of researchers' and participants' time and resources. It is in the best interest of application developers to create the strongest application possible before investing in a large-scale user study. The methods described in this paper are not as resource intensive, and can make the most of a future large-scale study by detecting problems earlier in the application development process.

Future work can be done to further enhance the mobile health application development process. To further refine the qualitative captology analysis and contextualize it within the field of mobile health technology, research could be done to create a framework for evaluation of persuasive capacity specifically of mobile health technology. This would partially address the limitation of the captology analysis where some specifications of SlipBuddy were undetectable by existing captology frameworks. Additionally, conducting a large-scale user study with SlipBuddy would help us generate further results with which we can compare the results of these small-scale methods. As mobile health continues to grow, efforts to make the development process more effective and efficient should grow as well. In this way, the field can better serve users in making better choices for their health.

Acknowledgements

This research was supported in part by the University of Massachusetts Science and Technology Initiatives Fund and the National Science Foundation (NSF) under Grant CNS-1560229 REU SITE: "Data Science Research for Safe, Sustainable and Healthy Communities." Any opinions, findings, conclusions, and/or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

References

1. Tulu B, Ruiz C, Allard J, et al. SlipBuddy: A Mobile Health Intervention to Prevent Overeating. Paper presented at: Proceedings of the 50th Hawaii International Conference on System Sciences2017.
2. Fogg BJ. Persuasive technology: using computers to change what we think and do. *Ubiquity*. 2002;2002(December):5.
3. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems.* 2009;24(1):28.
4. Petty RE, Cacioppo JT. The elaboration likelihood model of persuasion. *Advances in experimental social psychology.* 1986;19:123-205.
5. Tam KY, Ho SY. Web personalization as a persuasion strategy: An elaboration likelihood model perspective. *Information systems research.* 2005;16(3):271-291.
6. Manning C, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge University Press; 2008. Available from: https://nlp.stanford.edu/IR-book/
7. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Paper presented at: Acm sigmod record1993.
8. Agrawal R, Srikant R. Fast algorithms for mining association rules. Paper presented at: Proc. 20th int. conf. very large data bases, VLDB1994.