# Heuristic Evaluation of Persuasive Health Technologies

Julie A. Kientz[1,2], Eun Kyoung Choe[1], Brennen Birch[2], Robert Maharaj[2],
Amanda Fonville[1], Chelsey Glasson[2], Jen Mundt[2]
The Information School[1] and Human Centered Design & Engineering[2]
University of Washington
Seattle, Washington, USA
{jkientz,eunky,bpbirch,maharr,ajf32,chlsy}@uw.edu, jen.mundt@gmail.com

## ABSTRACT

Persuasive technologies for promoting physical fitness, good nutrition, and other healthy behaviors have been growing in popularity. Despite their appeal, the evaluation of these technologies remains a challenge and typically requires a fully functional prototype and long-term deployment. In this paper, we attempt to help bridge this gap by presenting a method for using heuristic evaluation to evaluate persuasive technologies. We developed a set of 10 heuristics intended to find problems in persuasive technologies that would affect persuasive elements, adoption, or long-term effectiveness of the technologies. We compared the performance of Nielsen's heuristics to our heuristics on two persuasive technologies using 10 different evaluators. Using our heuristics, evaluators found more severe problems more frequently. In addition, the issues that found only by our heuristics were more severe and more relevant to persuasive, cultural, and informational issues of the interfaces evaluated. Our method can be helpful in finding problems in persuasive technologies for promoting healthy behaviors earlier in the design process.

## Categories and Subject Descriptors

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous, K4.2: Computers and Society: Social Issues, J.3 Computer Applications: Life and Medical Sciences

## General Terms

Design, Human Factors

## Keywords

Persuasive technologies, heuristic evaluation, health, health informatics, heuristics, captology

## 1. INTRODUCTION

The use of persuasive techniques and behavior modification strategies for the design of technologies that promote healthy behaviors has been a major research trend of the past decade [10]. Researchers and industry both have developed a number of applications that use software, mobile technologies, games, and websites that encourage people to exercise more [5],[6],[16], eat more healthily [4],[8],[11], sleep better [14], or quit smoking [13],[22]. One of the challenges in designing for this space is in evaluating the effectiveness of these technologies, especially along the more persuasive aspects. Standard user interface evaluation techniques may be able to address some of the usability aspects, but are usually poor tools for assessing how successful a technology may be in helping to promote behavior change or the likelihood of long-term adoption. Thus, new ways of evaluating persuasive technologies for health have been a recent focus of a number of researchers in this space [9],[23].

Heuristic evaluation has been a classic and very popular evaluation method in the human-computer interaction community ever since it was first introduced by Jakob Nielsen in 1990 [21]. This discount usability technique enables a few expert evaluators to assess a technology artifact for problems that may affect its usability in a short amount of time along a set of guidelines known as heuristics. Practitioners have reported that heuristic evaluation is a good way to find significant usability problems at many stages of the design process [25]. Nielsen's technique has been widely adopted, and recently, there has been a trend to develop and validate more specialized heuristics for more specialized technologies. The intention of the specialized heuristics is to allow for evaluation of problems beyond usability and to focus evaluators on the topics that matter most to technology designers and the intent of the technology. Others have modified heuristics for technologies such as ambient displays [19], collaborative technologies [1], robots [3], and video games [24] and have been successful in identifying problems specific to those technologies, and in some cases, more severe problems than the traditional heuristics Nielsen describes [20].

Following this trend, we set out to develop and validate a set of heuristics designed to be more appropriate for evaluating technologies aimed to persuade users to make healthier choices. From our own experience of designing persuasive technologies, we have realized that although usability matters, it may have less of an impact on the persuasiveness of the system over other factors. Thus, we iteratively defined a set of 10 heuristics based on behavior change recommendations and guidelines from the persuasive technology literature and a modification of Nielsen's original heuristics. We designed a study that would allow us to validate our new heuristics by comparing them to Nielsen's original 10 heuristics to understand how they were used by evaluators. We also wanted to determine whether the new heuristics were able to find more errors, find more severe errors, or find errors of a specific type or those that are more relevant to persuasive technologies. We recruited 10 evaluators to conduct heuristic evaluations of two commercial products designed to promote behavior change: Mindbloom[1], a life goal setting website and MyFoodPyramid BlastOff[2], a game designed to teach children

*IHI'10*, November 11–12, 2010, Arlington, Virginia, USA.
Copyright 2010 ACM 978-1-4503-0030-8/10/11...$10.00.

[1] http://www.mindbloom.com

[2] http://www.mypyramid.gov/Kids/Kids_game.html

to make healthy food choices. Five evaluators used the new heuristics for the evaluation while the other five used Nielsen's original 10. We found that although our heuristics did not find more problems overall, they seemed to find more severe problems and problems that are more relevant to persuasive technologies.

The rest of this paper is organized as follows. We first provide a description of the related work in the areas of persuasive health technologies, evaluation of persuasive technologies, and heuristic evaluation. We next present the 10 new heuristics we developed along with a description of how they were determined. Following the heuristics, we describe the study we designed to compare the new heuristics to Nielsen's original 10 and then describe the results. We discuss the findings and conclude the paper.

## 2. RELATED WORK

In this section, we outline the related work that was most pertinent to this study. The major areas we focus on are in persuasive technologies for health, evaluation of persuasive technologies, and heuristic evaluation of novel technologies.

### 2.1 Persuasive Technologies for Health

In the past two decades, persuasive technologies have entered the marketplace at an increasingly accelerated rate. Although persuasive technologies serve many different purposes, some argue that their most significant contribution is in the domain of health [15]. Persuasive technologies for health currently span a variety of application areas. Some of the more popular ones include promoting physical fitness, such as through the Fish N' Steps game [16], the Nintendo Wii Fit[3], Kidnetic's games for kids[4], or Consolvo et al.'s Ubifit system [5]. Others have worked toward more general preventive health care, such as through healthy heart record-keeping[5], smoking cessation [13],[22], or preventing unwanted pregnancies[6]. Management of chronic disease is another area where cell phones [18] and websites are expanding our capabilities. There has even been some work in promoting personal hygiene, such as tooth brushing for children [2]. This list of application areas and persuasive technologies for health is by no means comprehensive but provides an insight into their vast domain. The research we conducted aimed to allow for the easier evaluation of technologies from any health domain.

Although websites may currently be the most common platform for persuasive technologies, this will not be the case in the near future [10]. In fact, health related persuasive technologies are already being developed for a variety of platforms including cell phones [5],[6], social networking sites[7,8], and video games[9,10]. In this work, we have created a method of evaluating the effectiveness of persuasive technologies regardless of their form factor, application type, or platform.

### 2.2 Evaluation of Persuasive Technologies

Persuasive technologies can be very difficult to evaluate for several reasons. The main reason concerns the fact that measuring the effectiveness of persuasive technologies requires a unique understanding of both the persuader and persuadee, in addition to how they interact with one another [23]. As a result of this complexity, many persuasive health technologies are evaluated through studies that require large budgets and a substantial amount of effort in addition to utilizing fully functional prototypes. These types of expensive studies are not always an option and sometimes overlook long-term persuasive effects. For this and other reasons, there is a need for the further development of evaluation techniques that can quickly and inexpensively detect usability problems early on in the design process of persuasive technologies for health, such as a specialized heuristic evaluation. Heuristic evaluation is a low-cost and accessible option that is already one of the top techniques currently used by usability practitioners [26]. Thus, we are seeking to improve upon the evaluation methods for use by designers of persuasive technologies for health, which may be used in determining their usability and likelihood of adoption, success, and persuasiveness.

### 2.3 Heuristic Evaluation

Jakob Nielsen created a method for using a set of guidelines to test the usability aspects of a product, which is known as heuristic evaluation. Heuristic evaluation is an informal method of usability testing that consists of a number of evaluators that are presented with an interface design, and then are asked to comment on the errors and effectiveness of the product [21]. In addition, heuristic evaluations are a low cost, fast, and efficient method of being able to identify any usability issues that may occur with the product.

Having people conduct these evaluations according to certain rules is the ideal method of using heuristic evaluation. Most people who use heuristic evaluations would perform them based on their intuition and common sense [21]. Currently, there are 10 recommended general usability heuristics that are based on Nielsen's methods of heuristic evaluation [20]:

- Visibility of system status
- Match between system and the real world
- User control and freedom
- Consistency and standards
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design
- Help users recognize, diagnose, and recover from errors
- Help and documentation

Heuristic evaluations can be applied to any interactive technological product. Many research studies have used heuristic evaluation on specific technologies, such as ambient displays, video games, collaborative technologies, and electronic medical records, as we describe below. The majority of the following research studies evaluated a set of new heuristics that they created and compared it to Nielsen's original set of heuristics, which served as a model for our study design.

The heuristic evaluation for ambient displays involved creating a modified set of heuristics and comparing them to Nielsen's original heuristics on two ambient displays. The results showed that the modified version was able to help the evaluators find more severe problems. In addition, the research demonstrated that using heuristic evaluation is an effective way of identifying the usability problems with ambient displays [19]. Video game designers have relied on using heuristic evaluations to determine the usability problems that might hinder their product's ability to be effective. A research study describes the importance of

[3] http://www.wiifit.com
[4] http://www.kidnetic.com
[5] http://www.iphonehealthapps.net/2009/11/05/Healthy-Heart-Journal/
[6] http://www.stayteen.org/
[7] http://www.healthchapter.com
[8] http://www.dailyburn.com
[9] http://easportsactive.com/
[10] http://www.respondesign.com/

interaction with video games, and how problems with usability could affect the user's ability to play [24]. The research group decided to create a new set of heuristics that would be able to inspect the usability aspects of video games. As a result, the 10 new usability heuristics described how to avoid problems that were commonly overlooked. Heuristic evaluations can also be used to evaluate collaborative technologies. One study focused on the concepts of groupware by developing discount evaluation methods that focused specifically on groupware usability problems. The results suggested that heuristic evaluation using the groupware heuristics can be an efficient method for identifying teamwork problems in shared workspace groupware systems [1]. Finally, heuristic evaluation has even been applied to the evaluation of human-robot interaction [3].

More closely related to the realm of health technologies, heuristic evaluation can be used to evaluate personal health record systems. The original heuristics and a modified version of heuristics were used on three personal health record systems. The results indicated that the original 10 heuristics found many usability errors; however, the modified version of the heuristics performed better with the same number of evaluators [17]. Our work differs in that we focus primarily on technologies designed to *persuade* the end user to be healthier, while personal health records are primarily for storage and access to medical records.

Overall, heuristic evaluation is a beneficial method that can help identify discrepancies in any product. However, certain results may vary depending on the product when using the original set of heuristic evaluations, so it might be necessary to create a modified version of heuristics to reveal more issues. In this work, we seek to replicate many of these previous studies but use a new domain which has not yet been explored.

# 3. HEURISTICS

The first step in this research was to determine a set of heuristics that we believe would better assess persuasive technologies designed to promote good health. In this section, we describe the process for defining and selecting the heuristics and then provide the list of our 10 persuasive health heuristics and their definitions.

## 3.1 Process for Defining Heuristics

Our approach was to review the literature and compile a master list of all usability guidelines and heuristics we could find relating to persuasive technologies. A group of 13 researchers and students with experience in user-centered design and usability (including the co-authors) each generated a list of 10 of what they believed were the most important guidelines and recommendations found from the literature (e.g., [7],[12]) and from their own experiences using and designing persuasive technologies. A total of 130 guidelines were generated. From this list, the researchers as a co-located group worked to narrow down the list by combining the similar guidelines, prioritizing them, and discussing them using a process similar to affinity diagramming. Several of the more specific guidelines were combined into one more general guideline, with the specifics included in the more descriptive text. In total, we chose to narrow the list of heuristics to 10 to keep a simple list that would not overwhelm evaluators and allow them to focus on the most important aspects. This also allowed us to do a comparison to Nielsen's original 10. We note that there is some overlap with Nielsen's original list, and this is intentional because in practice, this list would likely be used as a replacement for Nielsen's list and we did not want evaluators to neglect evaluating some of the more fundamental usability principles.

## 3.2 Persuasive Health Technology Heuristics

Below are the 10 final heuristics from the process we described. We gave each heuristic a short name and a longer definition.

1. **Appropriate Functionality:** The technology should meet usability, mobility, visibility, and durability needs according to the settings in which it might be used. The technology should function effectively in the user's environment by being easy to use and integrate into one's daily life and routine.
2. **Not Irritating or Embarrassing:** The technology should not irritate or embarrass the user, even after using the product repeatedly and regularly over a long period of time. This relates to aspects such as the presence of the product itself in the user's environment, the degree to which the technology intrudes upon the user's daily life, the timing, type, accuracy, and amount of feedback given, and the capability for customized settings and privacy controls.
3. **Protect Users' Privacy:** The system allows users to keep personal information private. Users can control what, when, to whom, how, and how much information is made public. Any public information is kept abstract.
4. **Use of Positive Motivation Strategies:** The technology recognizes when target behaviors have been performed or goals have been met and uses positive reinforcement strategies to promote continued progress. Avoids use of punishment for failure to perform target behaviors or meet goals.
5. **Usable and Aesthetically Appealing Design:** The visual design of the technology is attractive and appealing and adheres to basic usability standards. The design captures and sustains the user's interest, enhances user engagement with the technology, and also adds to the credibility and usability of the product.
6. **Accuracy of Information:** The technology should not inaccurately record or misrepresent the user's behavior (for instance, due to limitations in automatic sensing capabilities or the inability to use the device in certain environments). If necessary—to obtain an accurate, comprehensive account of behavior—the technology should allow users to edit data records and/or manually input additional data that the device is incapable of detecting automatically.
7. **Appropriate Time and Place:** Information, feedback, and assistance are provided at an opportune time and place (i.e., when and where it is needed, at the most appropriate time, and in the most effective manner).
8. **Visibility of User's Status:** The technology should always keep the user informed about progress toward goals through appropriate feedback within reasonable time. Feedback is accurate and easily understood (e.g., though use of abstract displays, summary data, etc.).
9. **Customizability:** Users should be able to customize aspects of the technology, for example, creating personalized goals and customizing product settings (public/private data, interface, etc.). However, customizability should not interfere with persuasive aspects.
10. **Educate Users:** Users should understand why the actions they do promote positive behaviors, and how their goals are being met. This includes which specific behaviors lead to the accomplishment of a larger goal. The technology should engage users in an active process whereby they learn information and gain skills relevant to their goals, particularly skills that would enable them to continue to progress towards goals even in the absence of the technology.

# 4. STUDY DESIGN

After developing our list of heuristics, we needed to validate that the new heuristics would be more successful at determining problems in design that may contribute to failure of a technology to promote behavior change. Thus, we designed a study that would compare the application of our new persuasive technology heuristics to Nielsen's original ten heuristics and their success at identifying potential problems in persuasive technologies. Our main goals with this study were to determine whether the new heuristics were understandable to evaluators and test the following hypotheses: the new heuristics will find 1) more severe issues, 2) more severe issues more frequently, and 3) more issues that are useful in improving persuasive aspects of the interface evaluated.

## 4.1 Technologies Evaluated

This section describes the two persuasive health technology products that we evaluated with heuristic evaluation. Both products are designed to promote behavior change in their users. As a group, we created a list of 13 readily available persuasive health applications that we could potentially evaluate. We then discussed the benefits of including each of the applications in our study and decided to choose two technologies that could be easily distributed to remote usability testers and covered two different persuasive approaches, as well as domains and level of specificity. In the end, we chose Mindbloom, a web-based application designed to track progress on your entire life's goals, including health, fitness, social relationships, and finance, and MyPyramid BlastOff!, a web-based game to teach kids about healthy food choices. We note that the specific choice of technologies is not crucial to the evaluation of our heuristics other than providing representative examples of persuasive technologies that covered a broad spectrum, and this study likely could have been replicated using different technologies.

### *4.1.1 Mindbloom*

Mindbloom is an online application that allows users to set short and long-term life goals and priorities and aims to build meaningful relationships between users. Mindbloom's primary users are life coaches and their clients. A user's life is represented by a "Life Tree" whose branches represent life areas important to each individual user, e.g. health, spirituality, relationships, leisure, lifestyle, finances, creativity, and career (see Figure 1). On a branch, each leaf represents a goal or dream related to that life area.

As a user takes steps to fulfill his or her goals, the Life Tree grows and the user is rewarded with seeds that can be used to unlock new features and grow their tree. Just as a user is rewarded with seeds for making progress on a goal, a user must spend seeds in order to set new goals or tasks. Mindbloom also incorporates a social networking aspect into the game. Users are encouraged to share their trees with friends; adding a social element to the game encourages them to visit regularly and keep up with their goals.

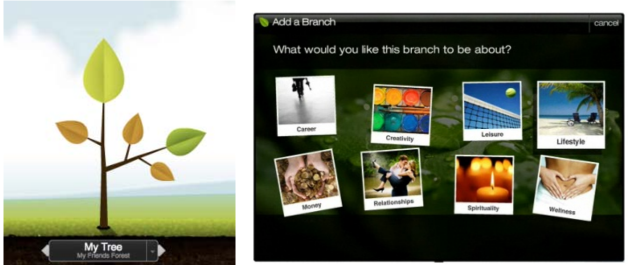

**Figure 1: Screenshots from Mindbloom.com. The left shows a tree with each branch representing a goal category and each leaf representing a specific goal. Colors indicate how recently the user has made progress toward the goal. The right shows the 8 categories of goals from which users can choose.**

### *4.1.2 MyPyramid Blast Off*

MyPyramid Blast Off is a game designed to educate children about the importance of healthy eating and physical activity. By demonstrating how children can select healthy foods for their own diets, the game persuades players to make smart choices about eating and exercise in their own lives. The game simulates a mission to space in which players must fuel their rocket ship and charge their battery in order to reach Planet Power. Players are instructed that the rocket requires one day's worth of fuel, represented by food, and sixty minutes of activity to launch (see Figure 2).

The game follows eating guidelines set by the United States Department of Agriculture; a player is given a recommended number of calories based on their age and gender and then selects food from the five categories of the Food Pyramid: grains, vegetables, fruits, milk, and meat and beans. Each food category is represented by a fuel gauge that fills up as a user adds food. Players are praised for adding healthy foods such as whole grains and 100% fruit juices. Additionally, players must log 60 minutes of physical activity to charge the rocket's battery for blast off. The rocket can only reach Planet Power if the user has made smart food choices and has performed 60 minutes of activity. The game is intended to help children explore food and exercise selections, rather than serve as a log of the day's activities.

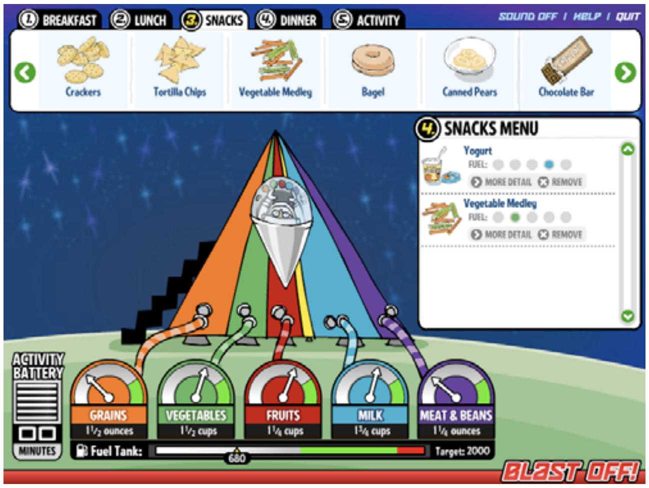

**Figure 2: Screenshot from the MyPyramid BlastOff! Game. Users "fuel" their rocket ship with different types of food from the different food groups, then try to launch their rocket. They win the game if they make healthy food selections, but lose if they do not.**

## 4.2 Participants and Recruitment

Expert usability evaluators participating in our study were recruited via word of mouth and through a number of online networks such as listservs and student or professional organizations (e.g., Usability Professionals' Association). Twenty-four individuals agreed to receive an email with information about the study, and 10 completed the study procedures. Among those 10 evaluators, eight were pursuing graduate-level degree in information management or HCI-related program, and the other two were a game designer and website coordinator. Evaluators' experience with user-centered design and usability ranged from less than 6 months to 6 years. Participants also provided self-rated experience with heuristic evaluations and persuasive technologies using a scale from 1 to 4 where 1 being "no experience" and 4 being "very experienced." They had some experience with heuristic evaluation ($M$ = 2.3) and slightly less experience specifically with persuasive technologies ($M$ = 1.78).

## 4.3 Study Protocol

In order to examine the effect of the persuasive health technology heuristics, evaluator participants were randomly assigned to two groups of five people—control and experimental group. We asked both groups to evaluate the two persuasive technologies (Mindbloom and MyPyramid) with the control group using Nielsen's heuristics and the experimental group using the new heuristics we developed from Section 3. Aside from the differences in heuristics, all individuals received identical instructions for conducting the evaluation and were given as much time as needed to do so. We instructed participants to spend approximately 15 minutes exploring each application before identifying as many usability and functionality issues as possible, using the heuristics as a guide. For each of the heuristics, participants wrote a short description of all the issues that fell into that category. Participants submitted their evaluations via email and received a $15 USD Amazon.com gift card upon completion.

After collecting all the data, the research group created a master list of issues based on the problems found by both the control and experimental groups. Because different evaluators had different scopes in defining a problem, the research group regrouped the problems then coded and mapped them onto the issues in the master list. For example, although the wordings of *two problems* found by different evaluators were not exactly the same, if they conveyed similar meanings, they were categorized into *one issue*. To maximize the validity of the severity rating, the members of the research group ($n$ = 10) individually rated the severities using Nielsen's recommendation [20] of using a 0 to 4 severity rating scale where 0 means "not a problem" and 4 means "usability catastrophe." The final severity rating was determined by taking the mode of all of the scores provided across the 10 researchers. To learn more about the attributes of issues found, the research group read through all the issues in the master list and created a code set with eight different types of usability problems using a bottom-up approach. Then, the members individually coded the issues using the code set. The frequency of issues found and which heuristic was used were also tracked. In the end, we created two master lists—one for each technology—which contain all the issues, frequency, severity ratings, and types of problems.

Unlike other studies where researchers first created a list of "known issues" and then identified the percentage of known issues found by the new heuristics [21], we created a master list based on the issues found by evaluators. Because we chose to evaluate more complicated systems which were beyond our control (e.g., they were websites we did not design ourselves), a list of known issues could never be comprehensive even with the input from more than 10 members from the research team. Thus, we believed examining the coverage of known issues was not the best way to assess the usefulness of the heuristics.

# 5. RESULTS

In this section, we describe the results of an analysis of the issues found by the evaluators and how they supported the research questions we aimed to answer in the design of the study.

## 5.1 Number of Issues Found

A total of 102 issues were found, 60 of which were found in Mindbloom and 42 in MyPyramid. The control group found 69 issues (44 from Mindbloom, 25 from MyPyramid) using Nielsen's heuristics while the experimental group found 62 issues (29 from Mindbloom, 33 from MyPyramid) using persuasive heuristics (Figure 3, Table 1).

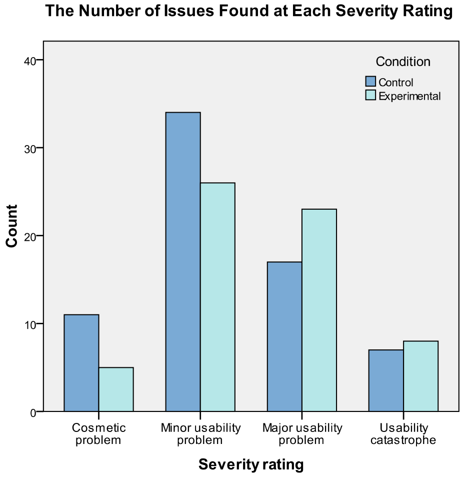

**Figure 3: The number of unique issues found at each severity rating. Both conditions did not find any issues at severity rating 0 – "not a problem."**

## 5.2 Issue Severity Hypothesis

Recall that our initial hypothesis was that the persuasive heuristics would be able to identify more severe issues than Nielsen's original 10 when evaluating persuasive technologies:

*Hypothesis 1: The issues found by the experimental group will be more severe than those found by the control group.*

To examine whether the persuasive heuristics found more severe issues than the Nielsen's heuristics, we compared the average severity ratings of the issues found in the two groups. The difference between the two groups in terms of average severity rating was marginally significant, with the issues found by the experimental group being more severe ($n$ = 62, $M$ = 2.55, $SD$ = .82) than those of the control group ($n$ = 69, $M$ = 2.29, $SD$ = .86), $t(129)$ = 1.75, $p$ = .082. However, when we removed the issues found in both groups and compare the remaining issues found only with Nielsen's heuristics ($n$ = 40) and with persuasive heuristics ($n$ = 33), the average severity rating of the issues found with the persuasive heuristics were more severe ($M$ = 2.45, $SD$ = .75) than those found with Nielsen's heuristics ($M$ = 2.03, $SD$ = .73), $t(71)$ = -2.46, $p$ = .016, indicating that the issues found by the experimental group were more severe than those found by the control group.

As shown in Table 1, when we examined the percentage within severity, the experimental group found more severe issues, such as the issues in severity ratings 3 and 4. When we examined the

**Table 1: The number of issues and percentages at each severity rating found by the control and experimental group. The last column shows the total issues found by either conditions. The highlights indicate which group found more of which type of issue.**

| | | | Condition | | Total Issues |
|---|---|---|---|---|---|
| | | | Control (Nielsen's) | Experimental (Persuasive) | |
| Severity Rating | 4 | Count | 7 | 8 | 9 |
| | | % within Severity | 77.8% | 88.9% | 100% |
| | | % within Condition | 10.1% | 12.9% | 8.8% |
| | 3 | Count | 17 | 23 | 31 |
| | | % within Severity | 54.8% | 74.2% | 100% |
| | | % within Condition | 24.6% | 37.1% | 30.4% |
| | 2 | Count | 34 | 26 | 48 |
| | | % within Severity | 70.8% | 54.2% | 100% |
| | | % within Condition | 49.3% | 41.9% | 47.1% |
| | 1 | Count | 11 | 5 | 14 |
| | | % within Severity | 78.6% | 35.7% | 100% |
| | | % within Condition | 15.9% | 8.06% | 13.7% |
| Total | | Count | 69 | 62 | 102 |
| | | % covered | 67.6% | 60.8% | 100% |

percentage within condition, the experimental group also found higher proportion of more severe issues.

## 5.3 Frequency of Severe Issues Hypothesis

Our second hypothesis in the design of our study was that the experimental group would find *more* severe issues:

*Hypothesis 2: The experimental group will find a higher number of more severe issues than those found by the control group.*

Regarding the number of total issues found, the experimental group did not find more issues than the control group (Table 1). Because not all issues are equally important, we were interested in knowing which set of heuristics was more likely to find more *severe issues*. In order to examine the effect of the heuristics on the ability to find more severe issues, we conducted a univariate analysis of variance test having two levels of condition (control, experimental) and five levels of severity rating (0 to 4) as factors, thus forming a 2 X 5 factorial design. The dependent variable was the average frequency of issues found. By average frequency, we mean the total number of issues found (frequency) divided by the unique number of issues at each severity rating (see Table 2); higher average frequency indicates that an issue is found more frequently more easily.

**Table 2: The number of issues, frequency, and average frequency at each severity rating found by the control and experimental group.**

| | | Condition | | | | | |
|---|---|---|---|---|---|---|---|
| | | Control (Nielsen's) | | | Experimental (Persuasive) | | |
| | | # of Unique Issues (A) | Frequency (B) | Average Frequency (B/A) | # of Unique Issues (A) | Frequency (B) | Average Frequency (B/A) |
| Severity Rating | 4 | 7 | 13 | 1.86 | 8 | 17 | 2.13 |
| | 3 | 17 | 30 | 1.76 | 23 | 38 | 1.65 |
| | 2 | 34 | 43 | 1.26 | 26 | 35 | 1.35 |
| | 1 | 11 | 13 | 1.18 | 5 | 7 | 1.40 |
| | 0 | 0 | 0 | N/A | 0 | 0 | N/A |
| Total | | 69 | 99 | 1.43 | 62 | 97 | 1.56 |

The omnibus F from the overall analysis revealed that there was a significant main effect of severity rating, $F$ (4, 10) = 14.07, $p$ < .001. To further investigate the main effect of severity rating, we conducted *a priori* contrast using the Dunn-Sidak correction. Evidence suggested that the average frequency of more severe issues, such as the severity rating 4—usability catastrophe ($M$ = 2.02, $SD$ = .50) and 3—major usability problems ($M$ = 1.73, $SD$ = .42) combined, was higher than the average frequency of less severe issues, such as severity rating 1—cosmetic problem ($M$ = 1.38, $SD$ = .48) and severity rating 2—minor usability problem ($M$ = 1.31, $SD$ = .07) combined, $F$ (1, 10) = 16.26, $p$ = .002. This indicates that severe issues were found more frequently in both conditions. The main effect of the condition and the interaction effect of severity and condition were not significant, although descriptive statistics from Table 1 and Table 2 and visual inspection of Figure 4 suggest that evaluators in the experimental group were more likely to find more severe issues more frequently using the persuasive health technology heuristics. In summary, although our heuristics did not find more issues, they found more severe issues more frequently.

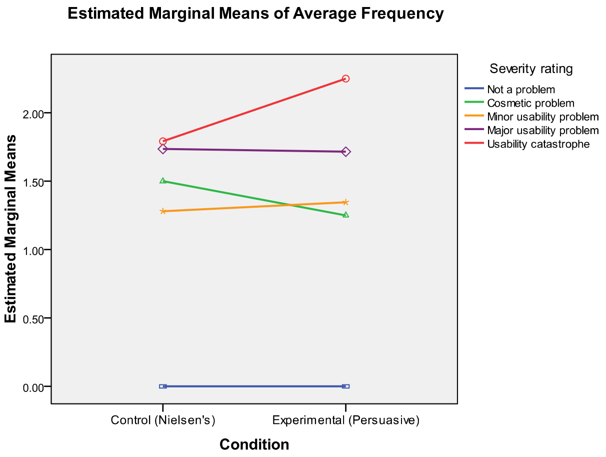

**Figure 4: Although the interaction effect was not statistically significant, the descriptive statistics indicate that the more severe issues (severity rating 4: usability catastrophe) were found more frequently in the experimental group using the persuasive health technology heuristics.**

## 5.4 More Useful Issues Hypothesis

Finally, our last hypothesis was that the new persuasive heuristics would find more useful issues than the control group:

*Hypothesis 3: The heuristics used by the experimental group will be more useful in evaluating persuasive technologies than those used by the control group.*

By *useful heuristics* for persuasive technologies, we mean the heuristics that are more sensitive to design problems related to the persuasiveness of the interface evaluated that may affect adoption rates or long-term use. To further investigate which of the two heuristics lists was more useful in evaluating persuasive technologies, we looked into the attributes of the issues that were found *only* in one condition or the other and categorized them using a code set consisting of eight types of issues or problems (see Table 3).

Figure 5 shows a breakdown of all the issues found ($N$ = 102). Each bar represents one issue, and the height of the bar indicates the frequency of the issue found. If an issue is found in both

**Table 3: A definition and example of each type of usability problem. All issues were categorized into eight different types of usability problems shown above. To maximize the validity of the categorizations, 10 people in the research team individually coded the issues and they were merged later.**

| TYPES | DEFINITION |
|---|---|
| AESTHETICS | Problems with the visual or auditory experience of the application, does not impede usability<br>*(e.g., The music is annoying)* |
| FUNCTIONAL | The application should do something conceptually but doesn't<br>*(e.g., The application should remember my login/password information)* |
| CLARITY | The application could be difficult to understand what to do<br>*(e.g., It's hard for me to know what to do first after the tutorial)* |
| PERSUASIVE | The application has issues that can affect its persuasive nature<br>*(e.g., The application doesn't sustain my interest over time)* |
| CULTURAL | The application does not always meet the diversity of its users<br>*(e.g., My character doesn't reflect my race or gender)* |
| NAVIGATION | The application has some problems in directing users through the site<br>*(e.g., I can't figure out how to get back to the home page)* |
| BUG | The application has a problem where it does not respond as expected<br>*(e.g., I entered my login information correctly, but nothing happens)* |
| INFORMATION | The application provides incorrect or ambiguous information<br>*(e.g., The application says I am unhealthy, but I have a normal BMI)* |

conditions, the blue and red bars are paired. If an issue is found only in either condition, there is only one bar, either in blue or in red, depending on in which condition it was found. Examining *Hypothesis 1*, we already learned that the issues found only with the persuasive heuristics (sole red bars in Figure 5) were more severe issues than those found only with Nielsen's heuristics (sole blue bars in Figure 5). With that in mind, we wanted to know how different they were in terms of issue type.

To investigate whether there was a relationship between the experimental condition and the type of issues found, we conducted a chi-square equality of proportion test. Evidence suggests that the experiment condition was related to the type of issues found, $\chi^2$ (7, $N$ = 73) = 15.40, $p$ = .03. Visual inspection of Figure 6 showed that the evaluators in the experimental group missed many functional and clarity problems which were found by those in the control group, whereas the evaluators in the control group missed many persuasive, navigation, information problems, and one cultural problem which was found by those in the experimental group. As described in Table 3, persuasive, cultural, and information problems are tightly related to the persuasive aspects of the interface evaluated that may affect the technology's adoption and long-term use. The attributes of the issue will be further discussed in the subsequent section.

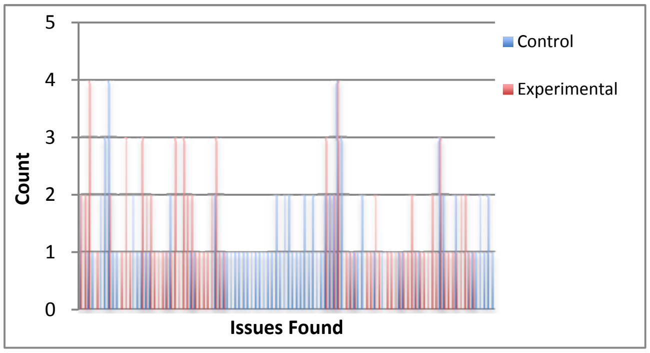

**Figure 5: A comparison of the coverage of the Nielsen's (blue) and Persuasive (red) heuristics for the Mindbloom and MyPyramid combined. Each bar represents one issue, and the height of the bar indicates the frequency of the issue found.**

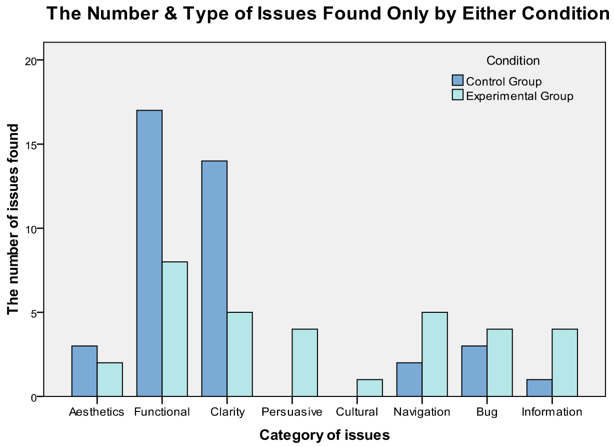

**Figure 6: Of the 73 issues that were found only in either condition, 40 issues were found with Nielsen's heuristics and 33 issues were found with persuasive heuristics. The researcher coded them using a code set that consists of the following eight categories of issue type: aesthetics, functional, clarity, persuasive, cultural, navigation, bug, and information.**

# 6. DISCUSSION

The results of our study indicate that the persuasive heuristics were effective in finding usability problems in the domain of persuasive health technologies. In comparison to Nielsen's original heuristics, we argue that a set of good persuasive health heuristics helps evaluators to find

- more severe issues,
- more severe issues more frequently, and
- more issues that are useful in improving persuasive aspects of the interface evaluated.

In this section, we discuss how heuristic evaluation as a design evaluation method can be improved for the domain of persuasive technology by reporting what we learned from this study.

## 6.1 Ordering of the Heuristics

We were interested in finding the most useful heuristic among the 10 persuasive heuristics that we gave to the evaluators. A useful heuristic would find more severe issues that are relevant to persuasive aspects of interfaces evaluated. The blue bar in Figure 7 indicates the number of issues found for each heuristic. It shows that heuristic #1, appropriate functionality, found the highest number of issues, 23, followed by heuristic #2, not irritating or embarrassing, which found 16 issues.

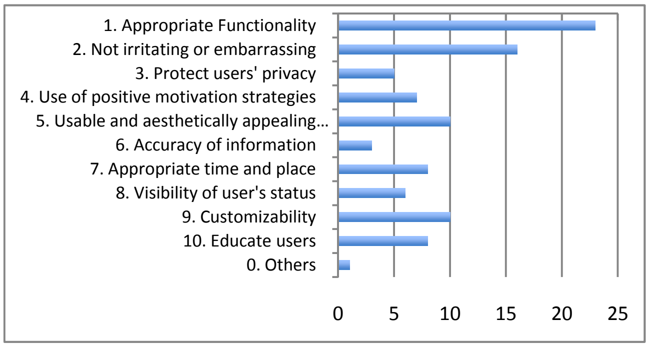

**Figure 7: List of persuasive heuristics and the number of issues found for each heuristic.**

However, we drew a similar bar graph using the frequency data from the control group using Nielsen's heuristics (Figure 8). Similar to Figure 7, the heuristic #1, visibility of system status, found the highest number of issues, followed by the heuristic #2.

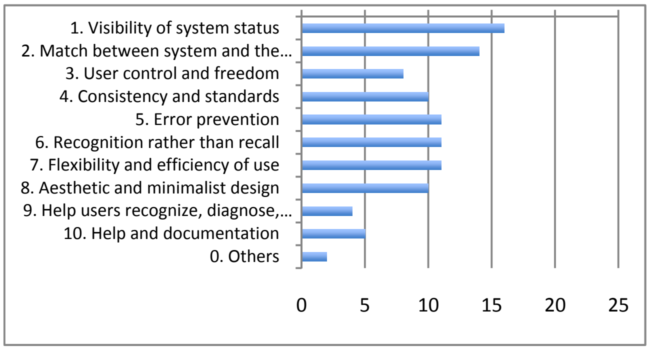

**Figure 8: List of Nielsen's heuristics and the number of issues found for each heuristic.**

When we created the 10 persuasive heuristics, we did not intend the ordering of the heuristics to be in the order of importance; however, it is possible for the evaluators of the system to think—either consciously or subconsciously—that the order of the heuristics matters, and thus try harder when applying the heuristics located at the front of the list. Therefore, we suggest ways to avoid the ordering effect by randomizing the order of heuristics for each evaluator when giving an instruction, or leveraging this by intentionally placing the heuristics in the order of importance. Further investigation of the ordering effect in a set of heuristics would help us in two ways: to gather more accurate heuristic evaluation data and to create more effective heuristics in the future.

## 6.2 Attributes of the Issues Found

Among the issues found, some issues were more relevant to persuasive aspects of an interface, such as supporting goal setting features, tracking progress, or providing helpful information and suggestions. These issues were difficult to find with Nielsen's heuristics and were also found less frequently even with the persuasive heuristics. However, they are important issues in the domain of persuasive technology. This is demonstrated in Figure 9, which shows that even if persuasive heuristic #1 found many issues, their average severity rating was lower than that of other heuristics, such as #2, #5, #6 and #10.

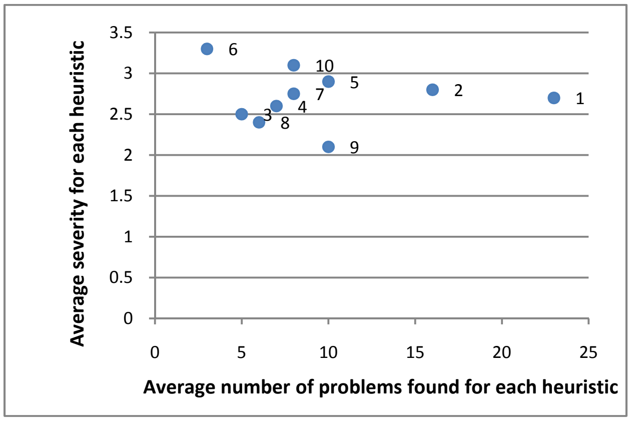

**Figure 9: A scatterplot of average number of issues found against average severity for each persuasive heuristic. The number next to each dot means the number of persuasive heuristics (#1 ~ #10).**

Thus, merely looking at the number of issues found with a heuristic does not tell us much about whether that particular heuristic is actually useful in finding persuasiveness-related issues. One success factor of our persuasive heuristics was that there was a general consensus among the research members to give higher severity ratings to the persuasive-related issues in evaluating persuasive technologies. For example, the statement below is one of the issues (severity rating: 3) which was found only by the experimental group twice using the persuasive heuristic #4, "use of positive motivation strategy":

> *"It is very black and white: you either pass or fail. Even when you have a completely nutritious meal, but you lack slightly in the veggie department, you fail completely. No kudos are given for the overall high quality meals that were constructed. Even exceeding by 5-10 calories causes failure, and the space shuttle does not launch. This lack of leeway/flexibly is causing the application to be less user-friendly and also the accuracy then becomes questionable."*

This is a great example of an issue that touches upon very important aspects of persuasive technology—use of positive motivation—but is very hard to find using Nielsen's heuristics. Useful persuasive heuristics help evaluators to find interesting issues that might prompt designers to think critically about the persuasive aspects of an interface. However, these types of issues are hard to find, and thus the number of issues found with a heuristic should not be read as the order of importance or usefulness.

## 6.3 Evaluating Long-Term Success

While heuristic evaluation is useful in finding many issues, there is still an issue of evaluating long-term success of persuasive technology, since some of the persuasive features will not be seen until after a certain amount of time has passed or only when certain contexts have been reached (e.g., unlocking of certain features after certain milestones are completed). For example, in the case of Mindbloom, the social aspects of the system required having friends in your social network rather than just using it alone. There was a demo video which showed these features, but interestingly, when some evaluators played a clip, they tended to point out the usability issues of the video player, but they did not necessarily pay attention to the contents of the video clip. We suggest that designers create demo accounts or pre-populated data for evaluation purposes so that evaluators can fully experience and understand all the persuasive features without having to use it for a long time. This may help expose some more issues related to long-term use of the different systems, which is a common goal for designers of health-based persuasive technologies.

## 7. CONCLUSIONS

While the evaluation of persuasive technologies for promoting health behavior change is still a difficult task, we hope the work we have conducted here will bring the research and design communities closer to attaining the goal of easy-to-evaluate persuasive technologies. With this research, we developed a list of 10 heuristics for persuasive health technologies that can serve as both guidelines for designers in the formative evaluation stages and as metrics for success in the summative evaluation stages. In a study we conducted to validate the heuristics, we had usability evaluators use the new persuasive heuristics to evaluate two different types of persuasive technologies: Mindbloom.com, a goal tracking website, and MyFoodPyramid Blast Off, a game designed to promote healthy eating in children. The evaluators were able to successfully conduct the study with the new heuristics, and when compared to evaluators using Nielsen's original 10 heuristics, they appeared to find more important issues. We found that the new heuristics, when compared to Nielsen's original 10 heuristics, were able to find more severe issues, find more severe issues more frequently, and find more issues of relevance to the persuasive nature of the technologies.

This work shows promise that discount usability techniques may be useful in the space of technologies designed to persuade users to engage in healthy behaviors. Although we do not recommend that heuristic evaluation be used as the sole evaluation technique for a persuasive technology, we hope that it will be useful to designers in finding issues and problems earlier in the design process, such as before they go through the resource-intensive process of building and deploying a fully functional prototype. In fact, we have already found that the problems identified by evaluators using our technique are useful. We shared the list of problems identified in this study with the Mindbloom.com usability designers. They were able to fix a number of the issues easily and refine the design to incorporate more persuasive elements. We hope that other practitioners can apply these heuristics broadly in their designs as well, not only for persuasive technologies for health, but in other areas of persuasive technologies beyond just health, such as sustainability, ethical behavior, or safety.

## 8. ACKNOWLEDGMENTS

We would like to thank our expert evaluators for participating in this study and Laura Barboza at Mindbloom.com for donating trial accounts for the evaluation. We also thank Sophia Fong, Steven Hudson, Linda Le, Dawn Sakaguchi, Kimberley Dietemann, Danyel Hacker, Heidi Sales, Kennen Pflughoeft, and Laura Schlenke for their assistance with this research. We thank Brian Landry and Jared Bauer for their feedback on the writing of this paper. This study was covered under University of Washington's Human Subjects Review Exemption #38217.

## 9. REFERENCES

[1] Baker, K., Greenberg, S., and Gutwin, C. 2002. Empirical development of a heuristic evaluation methodology for shared workspace groupware. In *Proceedings of the Conference on Computer Supported Cooperative Work CSCW '02*. ACM, New York, NY, 96-105.

[2] Chang, Y., Lo, J., Huang, C., Hsu, N., Chu, H., Wang, H., Chi, P., and Hsieh, Y. 2008. Playful toothbrush: ubicomp technology for teaching tooth brushing to kindergarten children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '08*. ACM, New York, NY, 363-372.

[3] Clarkson, E., Arkin, R.C.: Applying Heuristic Evaluation to Human-Robot Interaction Systems. FLAIRS Conference 2007: 44-49

[4] Connelly, K.H., Faber, A.M., Rogers, Y., Siek, K.A., Toscos, T.: Mobile applications that empower people to monitor their personal health. Elektrotechnik und Informationstechnik 123(4): 124-128. 2006.

[5] Consolvo, S., et al. 2008. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 1797-1806.

[6] Consolvo, S., Everitt, K., Smith, I., Landay, J.A. Design requirements for technologies that encourage physical activity, *Proceedings of the SIGCHI conference on Human Factors in computing systems*, April 22-27, 2006, Montréal, Québec, Canada.

[7] Consolvo, S., McDonald, D. W., and Landay, J. A. 2009. Theory-driven design strategies for technologies that support behavior change in everyday life. In *Proceedings of the 27th international Conference on Human Factors in Computing Systems*. CHI '09. ACM, New York, NY, 405-414.

[8] Denning, T., Andrew, A., Chaudhri, R., Hartung, C., Lester, J., Borriello, G., and Duncan, G. 2009. BALANCE: towards a usable pervasive wellness application with accurate activity inference. In *Proceedings of HotMobile '09*. ACM, New York, NY, 1-6.

[9] Fogg, B. 2009. Creating persuasive technologies: an eight-step design process. In *Proceedings of the 4th international Conference on Persuasive Technology. Persuasive '09*, vol. 350. ACM, New York, NY, 1-6.

[10] Fogg, B.J. Grudin, J., Nielsen, J., Card, S.. *Persuasive Technology: Using Computers to Change What We Think and Do*, Science & Technology Books, 2002.

[11] Grimes, A., Bednar, M., Bolter, J.D., Grinter, R.E.: EatWell: sharing nutrition-related memories in a low-income community. In *Proceedings of the 2008 Conference on Computer Supported Cooperative Work*: 87-96.

[12] Intille, S. S., "Ubiquitous Computing Technology for Just-in-Time Motivation of Behavior Change (Position Paper)," in *Proceedings of the UbiHealth Workshop*, 2003.

[13] Khaled, R., Fischer, R., Noble, J., Biddle, R. A Qualitative Study of Culture and Persuasion in a Smoking Cessation Game. *PERSUASIVE 2008*. pp. 224-236. 2008.

[14] Kim, S., Kientz, J. A., Patel, S. N., and Abowd, G. D. 2008. Are you sleeping?: sharing portrayed sleeping status within a social network. In Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work. CSCW '08. ACM, New York, NY, 619-628.

[15] King, P. and Tester, J. 1999. The landscape of persuasive technologies. Commun. ACM 42, 5 (May. 1999), 31-38.

[16] Lin, J.J., et al. "Fish'n'Steps: Encouraging Physical Activity with an Interactive Computer Game," *Proceedings of UbiComp '06*, (Sep 2006), 261-78.

[17] Liu, L.S. and Hayes, G.R. Heuristic Evaluation of Personal Health Records Systems. *Workshop on Interactive Systems for Health at CHI 2010*. 2010.

[18] Mamykina, L., Mynatt, E., Davidson, P., and Greenblatt, D. 2008. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the Conference on Human Factors in Computing Systems*. *CHI '08*. ACM, New York, NY, 477-486.

[19] Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., and Ames, M. 2003. Heuristic evaluation of ambient displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. ACM, New York, NY, 169-176.

[20] Nielsen, J. "Ten Usability Heuristics." useit.com. Retrieved 29 May 2010. http://www.useit.com/papers/heuristic/heuristic_list.html

[21] Nielsen, J. and Molich, R. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '90*. ACM, New York, NY, 249-256.

[22] Obermayer, J.L., Riley, W.T., Asif, O., Jean-Mary, J. 2004. College smoking-cessation using cell phone text messaging. *Journal of American College Health: J. of ACH*, 53(2), 71-78.

[23] Oinas-Kukkonen, H. and Harjumaa, M. 2008. A Systematic Framework for Designing and Evaluating Persuasive Systems. Persuasive 2008, 164-176.

[24] Pinelle, D., Wong, N., and Stach, T. 2008. Heuristic evaluation for games: usability principles for video game design. In *Proceedings the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. ACM, New York, NY, 1453-1462.

[25] Virzi, R.A., Sorce, J.F., Herbert, L.B. A Comparison of Three Usability Evaluation Methods: Heuristic, Think-Aloud, and Performance Testing. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, Computer Systems, pp. 309-313(5).

[26] Vredenburg, K. et al. A survey of user-centered design practice. In Proc. of CHI 2002, pp. 471–478. ACM Press, 2002.