Running Head: MYFROID.COM: A USABILITY ANALYSIS

myFroid.com:

A Usability Analysis Based on Cognitive Psychology Concepts

Alyssa Rosenzweig and Victoria Schwanda

University of Pennsylvania

Abstract

We are now in the age where computers play a significant role in our daily lives, so the interfaces through which we interact with them must be as easy to use as possible. Guiding principles from computer science, design and psychology inform the initial design of an interface, but they do not always come together seamlessly during the design process. Our experiment is an interface evaluation—also known as a usability study—of the web application myFroid.com. Using the think-aloud usability test method, participants were given seven tasks to evaluate the main features of the site and rated on their ability to complete the tasks. These performance ratings, as well as questionnaire feedback, from the first trial were used to make site alterations for the second trial. While our results are not statistically significant, they do show that participants performed better or at the same level in six of the seven tasks from trial one to trial two. Participants also gave higher ratings for all five ease-of-use questions in the post-study questionnaire in trial one than in trial two.

Introduction

We are now in the age where we depend upon and use computers on a daily basis. Since computers play such a significant role in our daily lives, the interfaces through which we interact with them must be as easy to use as possible in order to increase productivity. Beyond the increase in productivity, usable systems also reduce errors, reduce training and support, improve user acceptance of the system and result in an enhanced reputation for the developing organization (Maguire, 2001).Consequently, an entire field of study has developed that has impacted both industry and academia. This research field—human-computer interaction (HCI)—has emerged from a number of fields including computer science, design and psychology.

Computer scientists contribute to HCI by designing and building the systems and devices that we so heavily depend upon. Designers contribute to the field by applying design principles and theories to develop the interfaces for these systems and devices. Psychologists contribute to HCI by analyzing the target users and identifying principles from cognitive psychology that inform the overall system design.

Specifically, concepts such as meaningfulness, cognitive load and optimization, memory load, perception and cognitive maps are important contributions from cognitive psychology. To the field of HCI meaningfulness translates to the fact that people perform at their peak when they understand what they are doing. Therefore, interfaces should facilitate the user's understanding of the system so the user is better able to adapt to unexpected circumstances and solve problems. Systems should also be designed to minimize the amount of mental effort required to interact with it. When the system fails to minimize cognitive load, the result is reduced productivity and inability to focus on the task at hand. Interacting with interfaces is also an activity that places a

heavy load on working memory. As a result, important information must be repeated in multiple locations across the interface because users cannot be relied on to maintain important information throughout the interaction. The cognitive process of perception is facilitated by a meaningful organization of the visual elements of an interface. Design principles such as color, contrast and effective use of screen real estate enable the interface to be perceived as an organized entity. Similarly, a consistent interface that strategically utilizes labels and headers aids the user's development of a cognitive map of the system. This map is a factor that enables the user to successfully interact with the system.(Quesenbery, 1997)

While all of these principles should inform the initial design of an interface, they do not always come together seamlessly during the design process. Therefore, an interface evaluation—also known as usability study—is a critical next step in maximizing user productivity. While there are many different types of interfaces, including those found on cell phones, global positioning systems, heart monitors and computer software, a commonly tested interface is that of a web application. Our experiment is a usability study of a particular web application, myFroid.com.

myFroid.com is an online system where experimenters can create survey-based studies and participants can register and complete these studies. While myFroid.com was initially developed to create and host self-improvement studies, the system supports any kind of survey-based study. The use of the Internet as a medium to administer survey-based studies is a growing trend. One such organization utilizing this online method is the Penn Positive Psychology Center. The initial development of myFroid.com was inspired by a researcher involved with this organization.

A web usability study is usually conducted by an evaluator observing a user's interaction with the web application. There are many variations on methodology, one of which is the think-aloud method. The user is requested to perform certain tasks and verbalize their thoughts aloud to maximize feedback; the evaluator only intervenes if absolutely necessary (Maguire, 2001). For our study we conducted two trials, each with its own set of participants. After the first trial, using participant feedback and task performance, we made alterations to the site design and conducted the second trial using the altered design.

In each session the participant was asked to perform a series of seven tasks (detailed in Appendix A) using the think-aloud method. The seven tasks were chosen to simulate a complete interaction of the site from initial login and email registration to signing up for a survey and completing the survey to interactions that test additional features of the site embedded in the user's dashboard. The dashboard of the site is organized with a focus on four rounded-boxes (clockwise from upper left: To-Do List, Featured Experiment, Positive Psychology Feeds and How are you feeling today?) and to the left, a vertical navigation area including the participant's completed and available studies, as well as access back to the dashboard.

For the *Site Login* task, we wanted to evaluate if the participant could distinguish between the login and registration areas, as well as understand the overall concept of the website based on the brief description stated. The *Find Headlines* task highlighted the Positive Psychology Feeds box in the lower right corner, where recent updates from happiness and positive psychology blogs would rotate every few seconds. To determine whether participants would use the left navigation or the upper right Featured Experiment box, we asked each person to do the *Experiment Signup* task. Additionally, the *Experiment Signup* task entailed filling out a form with personal information (taking note of the effect of the asterisk as an indicator for a

required field), agreeing to the study terms and conditions, and then competing the first study survey. The *Complete Experiment* task led to another short survey and the *Find Response* task evaluated whether the participant could return to the study overview page to see a record of all their previously completed study activities and responses. In the How are you feeling? box in the lower left corner, when the user rolls the cursor over a face, a teal box forms around the face and the caption for the emotion appears below; clicking on a face will leave this border. To evaluate this interaction, participants were asked to *Record Happy* by selecting which face represented the happy emotion. For the *Find Emotion* task, the participant needed to recognize the linked text "View History" within that same box to reveal a graphical record of all their previously recorded emotions.

The presentation order of the tasks was randomized, taking into account the dependency between tasks such as *Record Happy* and *Find Emotion* because one had to be completed before the other. The importance of this ordering became particularly apparent in a pilot-run of the experiment where the task order was entirely random. This resulted in the participant having to do the illogical task sequence of *Find Emotion* before *Record Emotion*.

We created six presentation orders that corresponded to the six participant conditions (see Appendix B for all six task orders).Conditions were further put into one of three categories (A, B or C). They were categorized as follows: Condition A ordered the *Find Emotion* task directly after the *Record Emotion* task; Condition B ordered the *Complete Experiment* task directly after the *Experiment Signup* task; and Condition C ordered the *Find Response* task directly after the *Experiment Signup* or *Complete Experiment* tasks. We grouped conditions as such because we believed that in each grouping, the latter task would be easier in this ordering than in the orderings found in the other two conditions. This is because the user would end the first task on

the same screen as the screen where the second task would begin, which would cut down on the amount of navigation required. Also, since the tasks were related, the recent memory of the first task would facilitate the completion of the second task.

Based on the dependency relationships between certain tasks, we hypothesize that performance on individual tasks will vary by condition (*e.g. Find Emotion* task in Condition A1 versus other Conditions). We also hypothesize that the trial two participants will rank higher than the trial one participants on both the task performance evaluations and the ease-of-use questions in the post-study questionnaire. We also expect that by addressing the most common interaction difficulties listed in trial one (as measured by the three changes that each participant would make to myFroid.com) through the site modifications before trial two, these difficulties will not be listed in trial two.

Materials and Methods

Participants

Fourteen male and female undergraduate students from the University of Pennsylvania served as subjects in our study (7 male, 7 female).Due to technical difficulties, only six of the fourteen subjects (3 male, 3 female) were considered in the analysis. All subjects were between the ages of eighteen and twenty-one and reported using the Internet on a daily basis. Subjects received credit in introductory psychology courses for participating in our study.

Materials

Our experiment involved showing participants a website and asking them to complete seven tasks (see Appendix A for task descriptions) that were determined to be important in order to evaluate the site design and ease of interaction. While carrying out the task, the participant was asked to “think aloud” so the experimenter could take qualitative notes about the amount of

difficulty that the participant faced while completing the task. Based on the results from the first trial, some changes were incorporated into the site design and a second trial was run with different participants (see Figures 4, 5 and 6 for screenshots highlighting the differences between the two trials).

Although by title and description the site provided two different experiments available for participant registration in *Experiment Signup* task, both experiments asked the participant to answer the same two surveys. The goal of our experiment was to evaluate the website and interaction, so specific responses to the surveys presented on the site were not analyzed; participants were clearly informed of this to alleviate any privacy concerns.

Six conditions were assigned based on task presentation order, which was randomized taking into consideration certain dependent tasks that required other tasks to be completed prior to their completion (see Appendix B for task orders). The dependent relationships existed between tasks *Experiment Signup*, *Complete Experiment* and *Find Reponses* and between tasks *Record Happy* and *Find Emotions*. That is, task *Experiment Signup* had to occur before task *Complete Experiment* or *Find Response* and task *Record Happy* had to occur before task *Find Emotions*.

We used the *Firefox* Internet browser on PC-platform, desktop computers with a screen resolution of 1280x1024 pixels. All tasks were observed by the experimenter and manually rated according to participant successfulness. For some sessions, a trial version of the *My Screen Recorder* software was used to record cursor activity in a video format for later analysis.

Procedure

For the series of seven tasks, the experimenter stated the task and then gave the participant a slip of paper with that statement on it. The participant's thoughts that were stated

aloud, in combination with cursor activity, were evaluated to determine a task rating based on predetermined rating criteria (see Appendix C for rating criteria). After completing all seven tasks, participants responded to a six-question questionnaire. This questionnaire gathered basic demographic data, asked the participant to choose how much he or she agreed with a particular set of questions (see Appendix D for questions and rating scale), which translated into an average ease-of-use score for the participant, and asked the participants to list three things that they would change about the site. All participants were debriefed before the session ended.

The quantitative measure of interest was the rating of task scores and questionnaire responses, as well as the number of participants who cited a certain site attribute as something they would change. The qualitative measure of interest was experimenter observations and participant feedback on the post-study questionnaire. We looked at the cumulative average task scores of the first trial in comparison to the second trial and the average task score on tasks by condition between the first and second trials. Similarly, we looked at the difference between the cumulative average ease-of-use score from the post-study questionnaire between the two trials. We also looked at the number of occurrences of certain site attributes that participants would change between the two trials.

Results

The first quantitative measure of interest was the rating of task scores. We did not find a significant difference between the mean rating for any of the seven tasks across the three conditions between the two trials. All mean task rating scores can be found in Table 1.For the *Site Login* task, there was not a significant difference between the mean rating (M =4.33, SD = .47 for the all trial one conditions, M =5.00, SD = 0.00 for the all trial two conditions) across conditions between trials one and two ($t(4) = 1.69$, $p < 0.1$). For the *Find Headline* task, there

was not a significant difference between the mean rating (M =5.00, SD = 0.00 for the all trial one conditions, M =5.00, SD = 0.00 for the all trial two conditions) across conditions between trials one and two ($t(4) = 0.00$, $p < 0.1$). For the *Experimental Signup* task, there was not a significant difference between the mean rating (M =4.33, SD = 0.47 for the all trial one conditions, M =5.00, SD = 0.00 for the all trial two conditions) across conditions between trials one and two ($t(4) = 1.69$, $p < 0.1$). For the *Record Happy* task, there was not a significant difference between the mean rating (M =5.00, SD = 0.00 for the all trial one conditions, M =5.00, SD = 0.00 for the all trial two conditions) across conditions between trials one and two ($t(4) = 1.69$, $p < 0.1$). For the *Final Emotions* task, there was not a significant difference between the mean rating (M =3.33, SD = 0.47 for the all trial one conditions, M =4.67, SD = 0.47 for the all trial two conditions) across conditions between trials one and two ($t(4) = 2.39$, $p < 0.05$).For the *Complete Experiment* task, there was not a significant difference between the mean rating (M =4.67, SD = 0.47 for the all trial one conditions, M =5.00, SD = 0.00 for the all trial two conditions) across conditions between trials one and two ($t(4) = 0.84$, $p < 0.1$). For the *Find Response* task, there was not a significant difference between the mean rating (M =5.00, SD = 0.00 for the all trial one conditions, M =4.67, SD = 0.47 for the all trial two conditions) across conditions between trials one and two ($t(4) = 0.83$, $p < 0.1$).Additionally, there was not a significant difference between the mean of all the task means scores (M =4.52, SD = 0.66 for the all tasks across all trial one conditions, M =4.90, SD = 0.29 for the all trial two conditions) across conditions between trials one and two ($t(4) = 0.68$, $p < 0.1$).

The second quantitative measure of interest was the rating of post-study ease-of-use questionnaire response scores. We did not find a significant difference between the mean rating for any of the five questions across the three conditions between the two trials. All mean

questionnaire scores can also be found in Table 2. For the *myFroid.com is easy to use* question, there was not a significant difference between the mean rating (M =3.67, SD = .47 for the all trial one conditions, M =4.00, SD = 0.00 for the all trial two conditions) across conditions between trials one and two ($t(4) = 0.83$, $p < 0.1$). For the *I always know where I am in myFroid.com* question, there was not a significant difference between the mean rating (M =3.33, SD = 1.25 for the all trial one conditions, M =4.00, SD = 0.00 for the all trial two conditions) across conditions between trials one and two ($t(4) = 1.04$, $p < 0.1$). For the *It's easy to get lost in myFroid.com* question, there was not a significant difference between the mean rating (M =4.00, SD = 0.82 for the all trial one conditions, M =4.33, SD = 0.47 for the all trial two conditions) across conditions between trials one and two ($t(4) = 0.50$, $p < 0.1$). For the *myFroid.com is difficult to learn* question, there was not a significant difference between the mean rating (M =4.00, SD = 0.00 for the all trial one conditions, M =4.33, SD = 0.47 for the all trial two conditions) across conditions between trials one and two ($t(4) = .83$ $p < 0.1$). For the *I didn't get enough training* question, there was not a significant difference between the mean rating (M =3.67, SD = 0.47 for the all trial one conditions, M =4.00, SD = 0.82 for the all trial two conditions) across conditions between trials one and two ($t(4) = .50$, $p < 0.1$).Additionally, there was not a significant difference between the mean of all the question means scores (M =3.73, SD = 0.77 for the all tasks across all trial one conditions, M =4.13, SD = 0.50 for the all trial two conditions) across conditions between trials one and two ($t(4) = 0.68$, $p < 0.1$).

The qualitative measures of interest were experimenter observations (incorporated into task ratings) and participant feedback on the post-study ease-of-use questionnaire. Table 3 summarizes the number of participants who mentioned a specific type of site change.

Discussion

We found that our first hypothesis (task performance would vary by condition) was not confirmed by our data. We chose not to calculate inferential statistics on this data since we only had one participant per condition per trial (for a total of six participants), but instead looked at the individual task ratings and the standard deviations of the task ratings by task across conditions (still considering trial one and trial two separately).We found that all of the standard deviations were low and this, in addition to inspection of the data, allowed us to draw the conclusion that the conditions did not significantly influence task rating. After coming to this conclusion, we decided to collapse the conditions and consider the data by task as opposed to by condition for further analysis.

While this data contradicted our hypothesis, it is actually a positive result. We had originally hypothesized that task order (and therefore condition) would influence task performance because of the system flow involved in completing tasks. Different tasks require participants to navigate to different parts of the web application in order to successfully complete them. If the prior task left the participant on the same screen that the current task required the user to navigate to, there would be less navigation required to complete the current task. We believed that these types of interactions would result in easier or more challenging task sequences. The fact that we found this to not be case is a sign that the site is easily navigable.

Our second hypothesis was that participants would perform better on individual tasks as well as perform better overall in trial two compared to trial one. We did not find any significant results to support this hypothesis. While our results were not significant, we did see that in four of the seven tasks (*Site Login*, *Experiment Signup*, *Find Emotions*, *Complete Experiment*), the participants did perform better overall in trial two as we hoped the site alterations would do. In

two of the seven tasks, participants performed the same overall in trial one and trial two (*Find Headlines, Record Happy*).While all six participants received the same score for these two tasks, they all received the highest score possible indicating a high level of task success. In only one of the seven tasks did participant performance decline from trial one to trial two (*Find Response*).

We also expected to see a significant difference between the responses on the ease-of-use post-study questionnaire between trial one and trial two. While we did not find a statistical significance in mean question responses by question between trials, we did find that the mean question responses were higher for trial two than for trial one for each of the five questions. One potential explanation for this result is that participants' individual ratings are subjective. Participants were asked to rate the statements on a scale from *Strongly Agree* to *Strongly Disagree* and there is no way to make sure that participants are following a standardized scale when answering these questions. With a large sample size we could expect that these individualized differences would balance out, however this is less likely to happen with our small sample size.

We believe that the main reason for not finding statistical significance for our quantitative measures is our small sample size. With a sample size of only six participants, our results have limited statistical power. While our small sample size did affect the statistical analysis of our results, it is commonly said in usability literature that testing one subject is better than testing zero subjects (Krug, 2006). Every usability test yields useful data and will shed light on potential problems with the application's interface.

In addition to the quantitative data gathered through our experiments, we also gathered a large amount of qualitative data. In the post-study questionnaire participants were asked to state three changes that they would make to the system. We categorized these changes and then

counted how many participants mentioned these changes in trial one versus trial two (see Table 3). We found that participants in the first trial commonly cited changes that would help them understand how to use the website, such as clarifying the navigation, clarifying which information was required when signing up for experiments and providing feedback when certain actions are completed. Changes mentioned among participants in trial two tended to be more superficial such as making aesthetic changes, making the page more (or less) personalized and adding an *About Us* page.

These qualitative measures also helped to inform the changes that we made to myFroid.com between trial one and trial two. The first two changes were made to the website's dashboard and can be seen in Figure 4.The first change (marked by letter A) was to the left-hand navigation that provides a link to the dashboard as well as to studies that the participant can sign up for, studies that the participant is enrolled in and studies that the participant has completed. We renamed *Dashboard* to *Home* and made it the first link in the list as recommended by our users. We also renamed the label *Recruiting* to *Available* to more accurately denote studies that the participant could sign up for. These changes applied the cognitive map principle mentioned earlier. More accurate headings and heading placements help the user create a cognitive map and better understand the meaning of (and therefore how he or she can interact with) different portions of the site.

The second change was made to the *Positive Psychology Feeds* section of the site (marked with letter B). The second version of myFroid.com shows a fewer number of positive psychology headlines in order to better utilize whitespace as a means to separate content. It also utilizes different colors, sizes and font-weights to distinguish different headlines from the other

text. These changes are an example of the perception concept in action. By utilizing design principles we found a more successful way of organizing content.

The third change that we made was to the page that a user reaches when signing up for an experiment on myFroid.com. While signing up, certain information is required and certain information is not required. We found in our first trial that participants could not differentiate between these two categories. In an attempt to distinguish between these two categories, we positioned the asterisks further away from the labels (so they would be read separately and not as part of the label) and we also added a key to explain the meaning of the asterisks. This change is also an example of the perception principle because the added space between the label and the asterisk allowed for the user to realize that the asterisk was a symbol and had a greater meaning in the context of the interface.

The final change made to myFroid.com in between trials was also on the experiment signup page (see Figure 5). At the bottom of the page, users are required to check off boxes that they agree to the participant agreement (marked by letter A) and that they meet the participant requirements (marked by letter B) listed at the top of the page. In our first trial, users would commonly check off the first check box and assume that the second check box was the negation of the first (equivalent to I agree and I don't agree) whereas this second check box was actually to certify that they met the participant requirements. In the second version of myFroid.com we repeated the participant requirements in between the first and second check boxes so the user would understand that they were required to agree to two separate statements. This change applied the memory load principle mentioned earlier. While the participant agreement was at the top of the page, we could not count on the user's keeping this in mind as the completed the other required information. It is also an application of the perception principle because the participants

originally read the two check boxes as related to each other until they were separated by the text of the participant requirements.

Again, while we did not find that our changes led to a significant result for our hypothesis that performance would improve in trial two as compared to trial one, we did observe this trend. If we were able to run a third trial, we would incorporate additional changes such as making the left-hand navigation categories collapsible and adding more personalization through additional widgets and design features. These were some of the changes that were suggested by our trial two participants.

We also think that it would be interesting to include an eye-tracking component to our study. Much of our data came from experimenter observation of the participant's interaction with myFroid.com as well as the thoughts participants verbalized. We found that the *My Screen Recorder* software provided us with valuable information as to where the user was positioning his or her cursor and we believe that by adding an eye-tracking component to the study we would gain further insight into the process used by participants to navigate the interface.

It would also be of interest to test the interface with a non-college population where participant age, frequency of Internet use and computer fluency varied. We would also add a question to the post-study questionnaire about which computing platform (Mac or PC) the participant primarily used. We found that some terms such as "dashboard" had different meanings to certain participants who identified themselves as either Mac or PC users. Finally, we would also be interested in running similar usability tests on the experimenter side of myFroid.com. This side of the site is where experimenters log on and create the experiments that users can then sign up for. This process is already more demanding than the processes that the

non-experimenter users engage in, so we would be interested to see if our findings for the user side would be exaggerated by the already higher cognitive load placed on participants.

Lastly there are a few changes to our experimental design that we would make were we running this study again. We would first ensure that we could recruit more participants. Additional participants would have been extremely useful for the purpose of statistical analysis. Lastly, we would have made sure to ask the participants to tell us when they completed each task. While this may seem trivial, we found that there may have been times when we, as experimenters, realized that the participant completed the task and moved them on to the next task before they were ready. We might have collected additional data as to which parts of myFroid.com do not produce adequate feedback after task completion were we to wait until the participant is sure that he or she had completed the task.

References

Alshamari, M. and Mayhew P. (2008). Task design: its impact on usability testing. In *Proceedings of the Third International Conference on Internet and Web Applications and Services* (Athens, Greece). 583-589.

Chalmers P.A. (2003). The role of cognitive theory in human-computer interface. *Computers in Human Behavior*, 19, 593-607.

Chen S., Magoulas, G.D. and Macredie R. (2004). Cognitive styles and users' responses to structured information representation. *International Journal of Digital Libraries*, 4, 93-107.

Ford, N. (2004). Modeling cognitive processes in information seeking: from popper to pask. *Journal of the American Society for Information Science and Technology*.55, 769-782.

Hertzum, M., Hansen, K.D. and Andersen, H.K. (2009). Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload?. *Behaviour & Information Technology*, 28, 165-181.

Hornof, A.J. and Halverson, T. (2003). Cognitive strategies and eye movements for searching hierarchical computer displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA).249-256.

Jasperts, M.W.M., Steen, T., van den Bos, C. and Geenen M. (2004). The think aloud method: a guide to user interface design. *International Journal of Medical Informatics*, 73, 781-795.

Kaptelinin, V. (1995). Context and consciousness: activity theory and human-computer interaction. Cambridge, MA: Massachusetts Institute of Technology.

Krug, Steve (2006).Don't make me think. Berkeley, CA: New Riders.

Maguire, M.(2001). Methods to support human-centred design. *Int.J.Hum.-Comput.Stud.*, 55, 587-634.

Olson, G. and Olson J. (2003). Human-computer interaction: psychological aspects of the human use of computing. *Annual Review of Psychology*. 54, 491-516.

Quesenbery, W. (1997). Building documentation into the interface: a cognitive theory. In *Proceedings to 44th Annual STC International Conference* (Toronto, Ontario, CA). 249-256.

Sharp, H., Rogers, Y. and Preece, J. (2007).Interaction design: beyond human-computer interaction. West Sussex, England: John Wiley & Sons, Inc.

van Oostendorm, H. and Juvina, I. (2007). Using a cognitive model to generate web navigation support. *Int .J. Hum.–Comput. Stud.*, 65, 887-897.

Vaughan, M.W. (1998). Testing the boundaries of two user-centered design principles: metaphors and memory load. *International Journal of Human-Computer Interaction*, 10, 256-282.

Wroblewski, L. (2002). Site-seeing: a visual approach to web usability. New York, NY: Hungry Minds, Inc.

Appendix A

Task Names and Task Descriptions

Task Name	Task Description
Site Login	Please create an account and log in to myFroid.com.
Find Headlines	Please find three Positive Psychology headlines.
Experiment Signup	Please register for an experiment. Complete the first task of the experiment.
Record Happy	There is a way to record your emotions on this site; happy is one of them. Record this emotion using the tools on myFroid.com.
Complete Experiment	Complete the second task of the experiment that you already registered for.
Find Response	Now that you have completed a portion of the online experiment, please find your response for question #3 of the Baseline Questionnaire.
Find Emotions	Please show me the percentage of positive emotions you have recorded on myFroid.com.

Appendix B

Task Orders by Condition

	Find Emotions directly after Record Happy (A)		Complete Experiment directly after Experiment Signup (B)		Find Response directly after an Experiment Task (C)	
	A1	**A2**	**B1**	**B2**	**C1**	**C2**
Tasks (in order of presentation)	Site Login	Site Login	Site Login	Site Login	Site Login	Site Login
	Find Headlines	Experiment Signup	Record Happy	Record Happy	First Event	Find Headlines
	Experiment Signup	Record Happy	Experiment Signup	Find Headlines	Find Headlines	Record Happy
	Record Happy	Find Emotions	Complete Experiment	Experiment Signup	Record Happy	Experiment Signup
	Find Emotions	Find Response	Find Emotions	Complete Experiment	Complete Experiment	Find Response
	Complete Experiment	Find Headlines	Find Response	Find Emotions	Find Response	Complete Experiment
	Find Response	Complete Experiment	Find Headlines	Find Response	Find Emotions	Find Emotions

Appendix C

Rating Scale for Task Successfulness

Rating	Description
5	Participant completes task without prompting.
4	Participant completes task without prompting but with some difficulty.
3	Participant requires prompting to complete task.
2	Participant requires prompting and still completes task with some difficulty.
1	Participant fails to complete task.

Appendix D

Rating Scale for Questionnaire Responses

Question	Response					
	Strongly Agree	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**	**Don't Know**
myFroid.com is easy to use	5	4	3	2	1	0
I always know where I am in myFroid.com	5	4	3	2	1	0
It's easy to get lost in myFroid.com	1	2	3	4	5	0
myFroid.com is difficult to learn	1	2	3	4	5	0
I didn't get enough training	1	2	3	4	5	0

Table 1

Task rating by task and trial

Task	B1 Condition		B2 Condition		C2 Condition		Mean		SD	
	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2
Site Login	4	5	4	5	5	5	4.33	5.00	0.47	0.00
Find Headlines	5	5	5	5	5	5	5.00	5.00	0.00	0.00
Experiment Signup	4	5	4	5	5	5	4.33	5.00	0.47	0.00
Record Happy	5	5	5	5	5	5	5.00	5.00	0.00	0.00
Find Emotions	4	5	3	4	3	5	3.33	4.67	0.47	0.47
Complete Experiment	5	5	4	5	5	5	4.67	5.00	0.47	0.00
Find Response	5	4	5	5	5	5	5.00	4.67	0.00	0.47
Mean	4.57	4.86	4.29	4.86	4.71	5.00	4.52	4.90		
SD	0.49	0.35	0.70	0.35	0.70	0.00	0.66	0.29		

Figure Caption

Figure 1. Mean task ratings by trial.

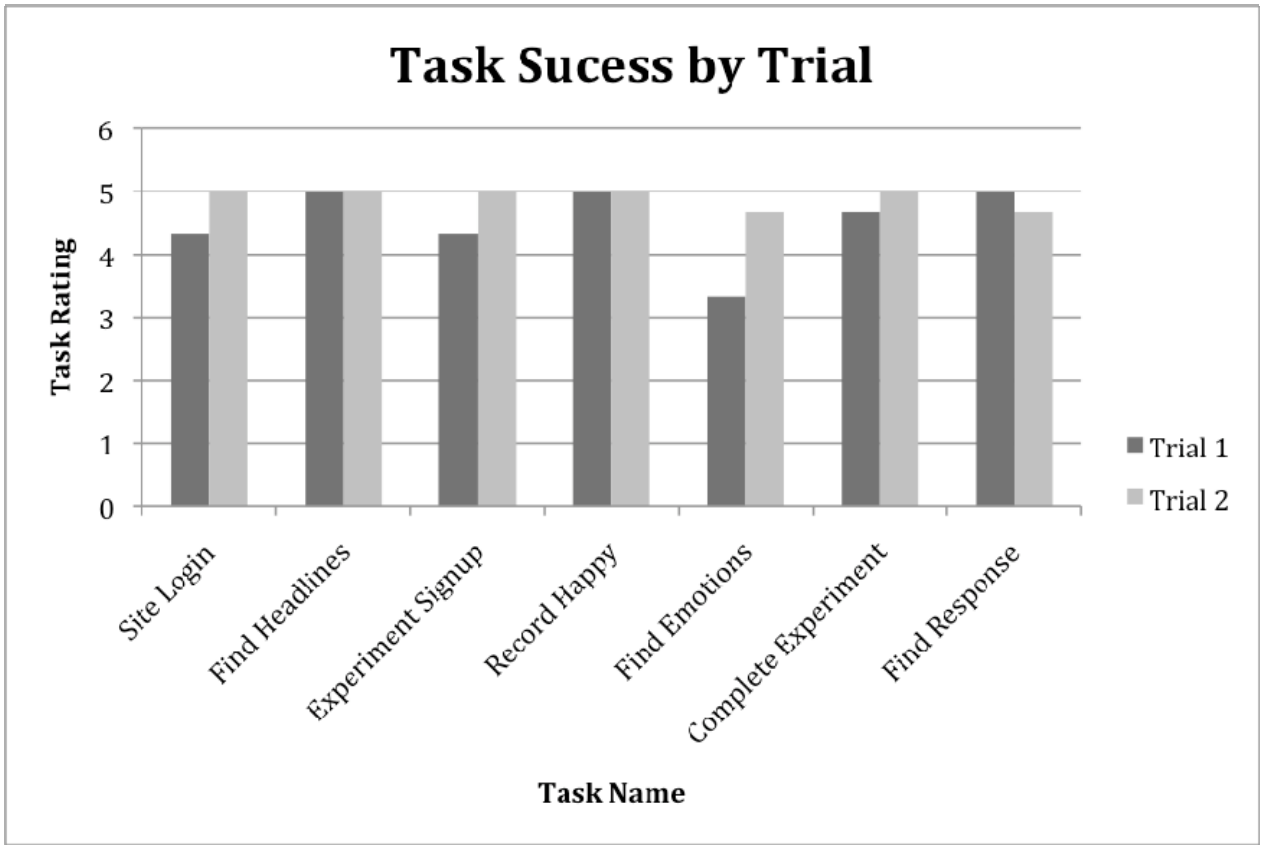

Table 2

Post-survey ease-of-use questionnaire responses by task and trial

Question	B1 Condition		B2 Condition		C2 Condition		Mean		SD	
	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2
myFroid.com is easy to use	3	4	4	4	4	4	3.67	4.00	0.47	0.00
I always know where I am in myFroid.com	2	4	5	4	3	4	3.33	4.00	1.25	0.00
It's easy to get lost in myFroid.com	3	4	5	4	4	5	4.00	4.33	0.82	0.47
myFroid.com is difficult to learn	4	4	4	4	4	5	4.00	4.33	0.00	0.47
I didn't get enough training	3	4	4	3	4	5	3.67	4.00	0.47	0.82
Mean	3.00	4.00	4.40	3.80	3.80	4.60	3.73	4.13		
SD	0.63	0.00	0.49	0.40	0.40	0.49	0.77	0.50		

Figure Caption

*Figure 2.*Mean post-study ease-of-use questionnaire responses by trial.

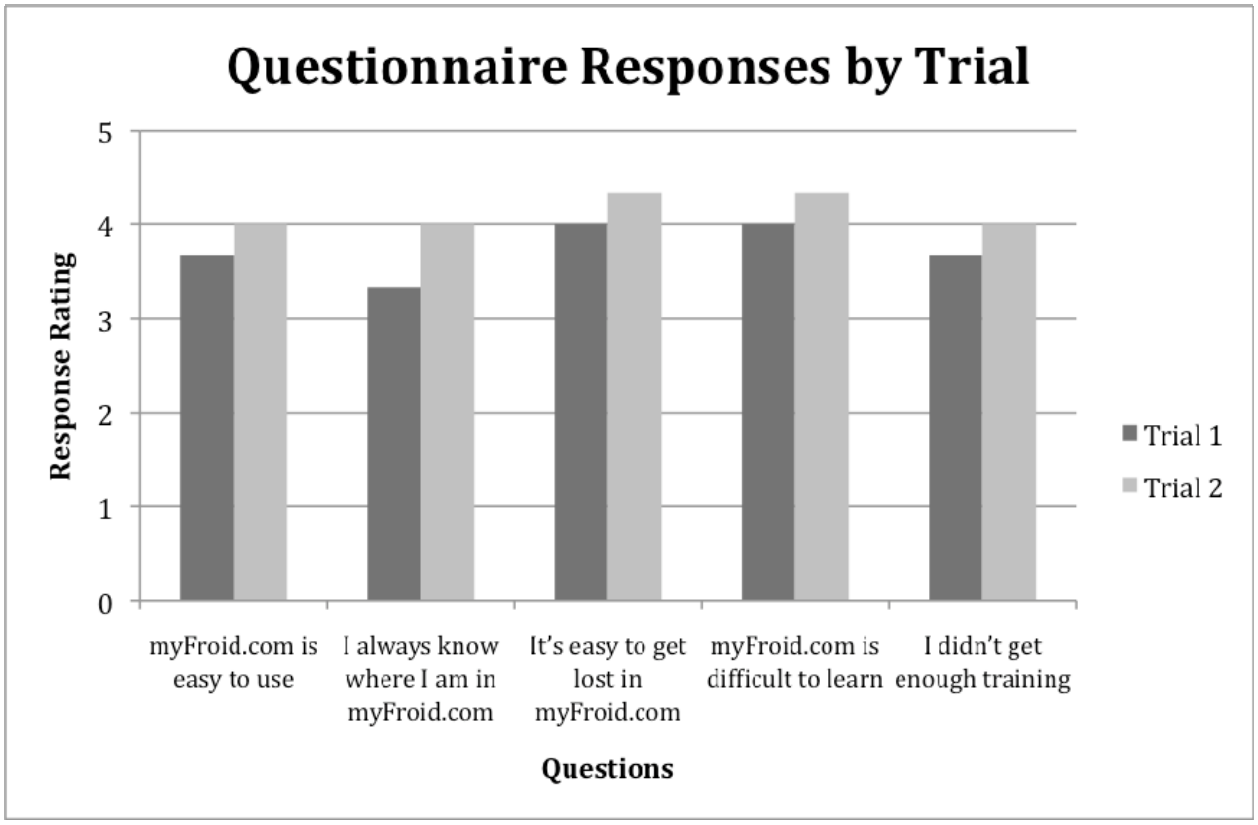

Table 3

Number of participants mentioning particular changes by trial

Change	Trial 1	Trial 2	Trial 1 – Trial 2
Changes to side navigation	5	2	3
Make required information clearer	1	0	1
Give feedback when registering for site	1	0	1
Add breadcrumb navigation	1	0	1
Aesthetic changes	0	1	-1
Add additional information about site	0	2	-2
Make home page more neutral	0	1	-1
Make home page more personal	0	1	-1
Prefer list of experiments to Featured Experiment	1	1	0

Figure Caption

Figure 3. Number of participants that mentioned particular changes in their post-study questionnaire by trial.

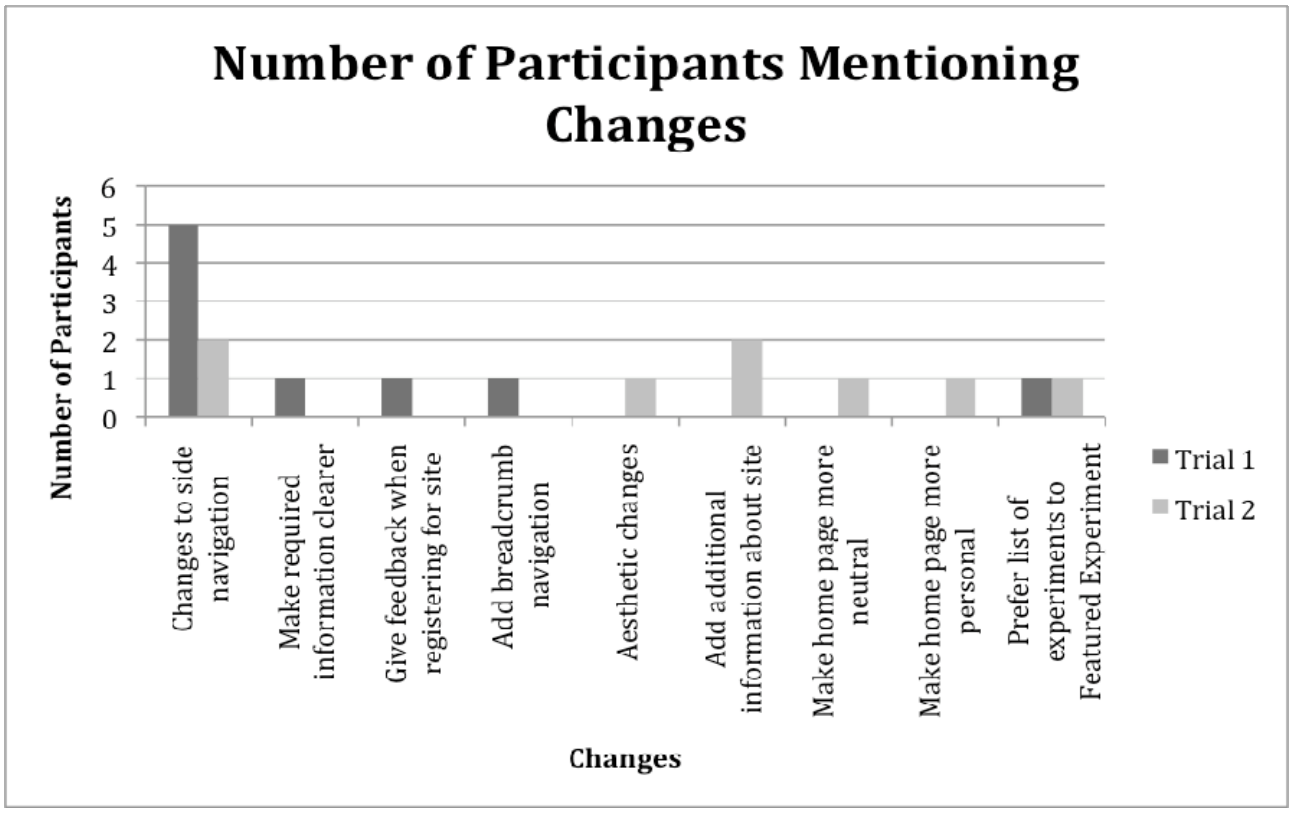

Figure Caption

*Figure 4.*Screenshots of myFroid.com's dashboard. The first image is a screenshot of the website used for trial one and the second image is a screenshot of the website used for trial two. The letters A and B call out the differences.

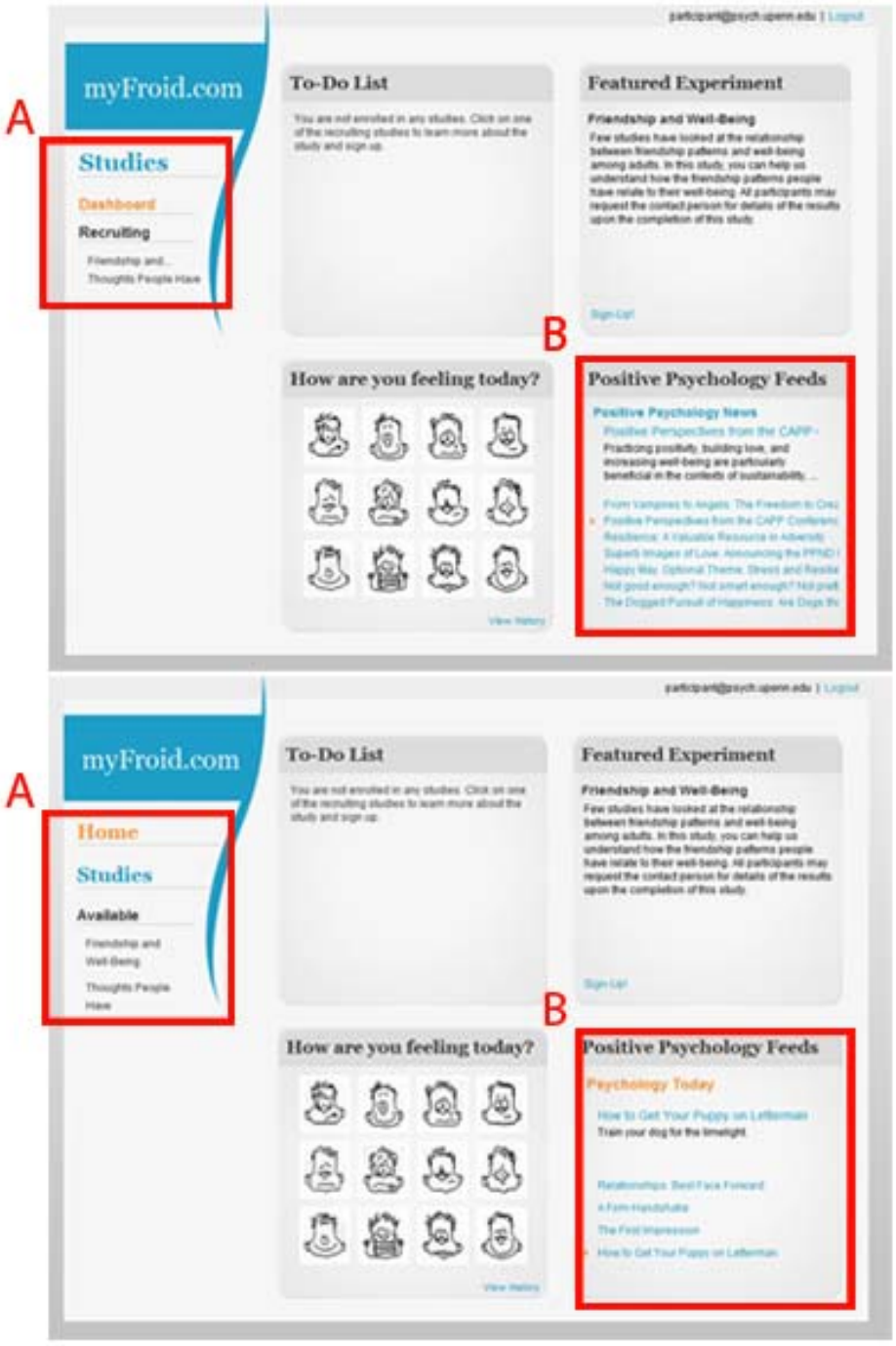

Figure Caption

Figure 5.Screenshots of the top half of myFroid.com's experiment signup page. The first image is a screenshot of the website used for trial one and the second image is a screenshot of the website used for trial two. The letters A and B call out the differences.

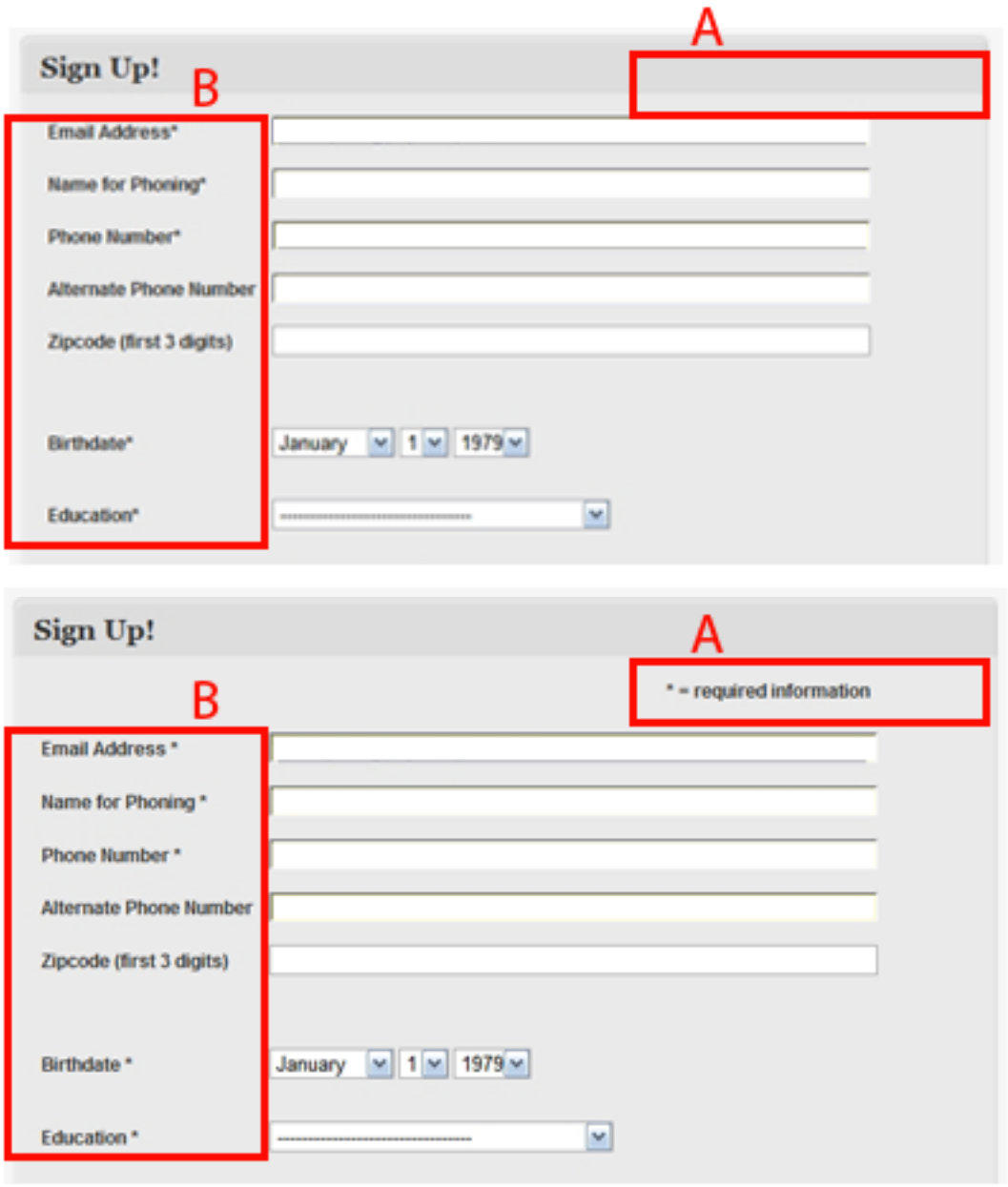

Figure Caption

*Figure 6.*Screenshots of the bottom half of myFroid.com's experiment signup page. The first image is a screenshot of the website used for trial one and the second image is a screenshot of the website used for trial two. The letters A and B call out the differences.