CHAPTER

Case Studies 10

In this chapter we present six case studies showing how other usability researchers and practitioners have used metrics in their work. We would like to thank the authors of these case studies: Hoa Loranger; Jim Lewis; Bob Bailey, Cari Wolfson, and Janice Nall; Scott Weiss and Chris Whitby; Agnieszka (Aga) Bojko; and Todd Zazelenchuk.

10.1 REDESIGNING A WEBSITE CHEAPLY AND QUICKLY—By Hoa Loranger

A well-known company in the entertainment news and content industry asked me to help with their website redesign project. The main goal of the redesign was to increase the number of site visitors and motivate them to spend more time exploring the site's content and offerings. Since much of the revenue generated on the site was from paid advertising, the number of page views and ad impressions was vital to the success of the website.

The challenge was to create a new site that was appealing to visitors while supporting business needs. If ads needed to stay, then the team needed to develop advertising strategies that were acceptable and, even better, appealing to users.

Similar to other redesign projects I have worked on in the past, this one needed to be completed quickly and cheaply. We had four months for multiple iterations of design and usability testing. The best way to tackle this challenge was to do it incrementally. Instead of trying to design the perfect solution at once, which is virtually impossible, we took a multiphase approach.

10.1.1 Phase 1: Testing Competitor Websites

Before delving too deeply into redesign, we first learned from the successes and failures of other similar websites by conducting a competitive study. The first phase was designed to discover people's reaction to concepts of the new site and to find out what worked and what did not work on a few other competitor sites. In a one-on-one lab environment, I gave participants tasks to do on three different websites and observed their behavior as they attempted to accomplish the tasks. Fifteen people participated in this study. All participants were interested in movies and entertainment and had a least a year of web experience, although most had significantly more.

Sample Tasks

Your friend just told you about a movie that is called [*Movie name*]. Check out *www.companyname.com* to see whether you might be interested in seeing this movie.

- Find the most convenient place and time to see [*Movie name*].
- Find out what TV shows and movies [*Actor name*] has been in.

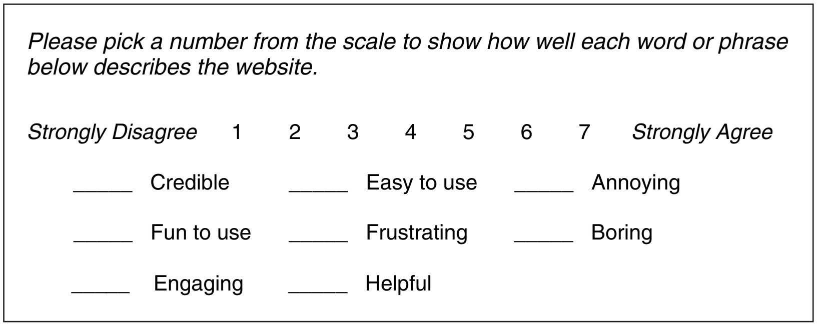
It was important that I stressed to the team that although this part of the study was termed "Competitive," it was not meant to be truly competitive. I reiterated (and reiterated) that we were not interested in seeing what websites people picked as the "winner." The purpose of this test was to identify the elements of each design that worked well and elements that did not work so well, so that we could eventually produce the best single design with the best aspects of all of them.

Questionnaire

After each site evaluation, participants were asked to rate the site on different characteristics. (Ratings were not used on the Company X design because it was still in its infancy and not comparable to the competitive sites.) The questionnaire is shown in Figure 10.1.

Results

The scores from the questionnaire were combined, resulting in the Website Appeal score. The average Website Appeal results are shown in Figure 10.2. This



Please pick a number from the scale to show how well each word or phrase below describes the website.

Strongly Disagree 1 2 3 4 5 6 7 *Strongly Agree*

_____ Credible _____ Easy to use _____ Annoying

_____ Fun to use _____ Frustrating _____ Boring

_____ Engaging _____ Helpful

FIGURE 10.1

Questions used in Phase 1 to measure website appeal.

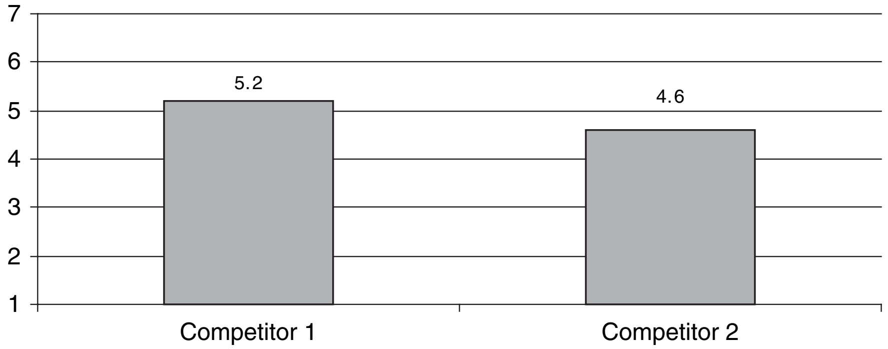


FIGURE 10.2

Average score for website appeal. Competitor 1 scored higher in desirable attributes than Competitor 2.

study helped the team identify major usability issues with the conceptual design, as well as provided insight into the advantages and disadvantages between other competitive sites.

10.1.2 Phase 2: Testing Three Different Design Concepts

Seven people participated in this part of the study. Based on the qualitative and quantitative information learned from Phase 1, the designers created alternative visual designs for the new website.

We did not have the time or budget to have different studies to measure graphical design and advertising conditions separately. We made the most of what we had by combining the various conditions. Again, we were relying on both quantitative data (people's comments and reactions) and the survey data to point us in the right direction.

The Designs Tested

Each participant evaluated two out of the following three designs.

- *Version 1*: Medium graphical treatment with high level of unrelated ads
- *Version 2*: Minimal graphical treatment with low level of related and unrelated ads
- *Version 3*: High graphical treatment with moderate level of related ads

I first asked them to provide their first impressions of the homepage and then gave them tasks to do on the websites. I noted user interactions, comments, and success scores for each task. After attempting tasks on each design, participants filled out a questionnaire to measure their preference.

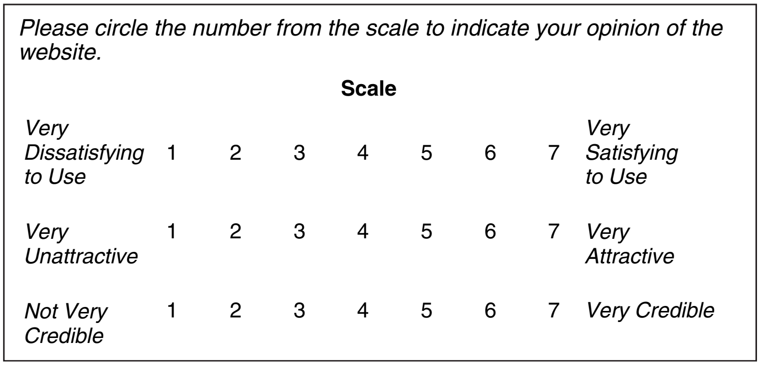
Questionnaire

After each site evaluation, participants were asked to rate the site. The questionnaire is shown in Figure 10.3.

Results: Satisfaction

In general, respondents were relatively satisfied with all three designs. When comparing the overall satisfaction scores of the three designs, there was only a moderate difference between people's satisfaction ratings for each design (see Figure 10.4). For example, between Design 2 and Design 3, there is only approximately a half-point incremental difference.

From a usability standpoint, the prototypes performed virtually the same. Most people were able to accomplish the tasks without much difficulty. This was not surprising because the interaction design for each version was virtually identical.



Please circle the number from the scale to indicate your opinion of the website.

Scale

Very Dissatisfying to Use	1	2	3	4	5	6	7	*Very Satisfying to Use*
Very Unattractive	1	2	3	4	5	6	7	*Very Attractive*
Not Very Credible	1	2	3	4	5	6	7	*Very Credible*

FIGURE 10.3

Questions used to measure design preference.

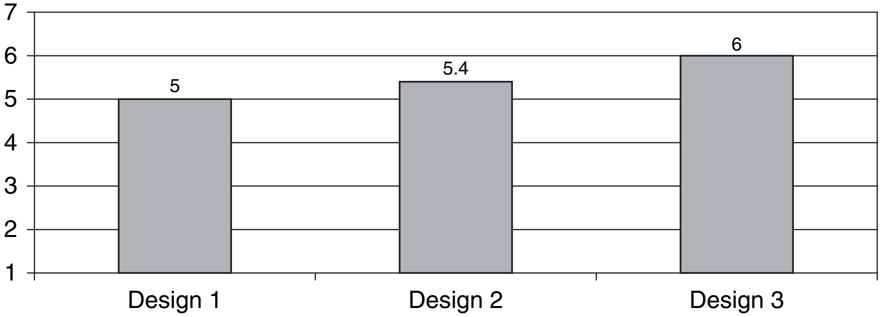


FIGURE 10.4

Phase 2: Satisfaction scores of three different website designs.

Therefore, differences in satisfaction ratings could mainly be attributed to the visual design differences and/or advertising conditions.

Results: Attractiveness and Credibility

When people scored the different sites strictly on attractiveness and visual appeal, Design 3 scored much higher than the other two designs, and they consistently chose Design 3 as their preferred choice (six out of seven users). In contrast to the results of the satisfaction ratings, there was a much greater difference in attractiveness scores (see Figure 10.5). Design 3 was rated much higher in attractiveness than the other two designs. The results from the attractiveness rating are consistent with what people said they would pick as their favorite design when given the option to choose one of the three visual designs.

Design 3 had the most attractive design, which also scored highest in credibility, even though it contained a moderate level of related ads (see Figure 10.6). People

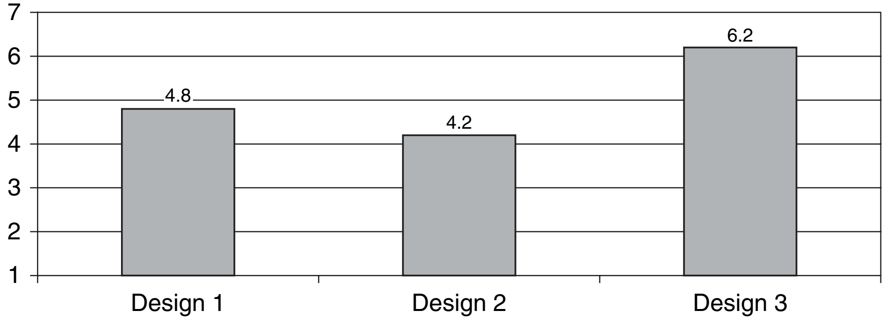


FIGURE 10.5

Phase 2: Attractiveness scores of three different website designs. Six out of seven participants chose Design 3.

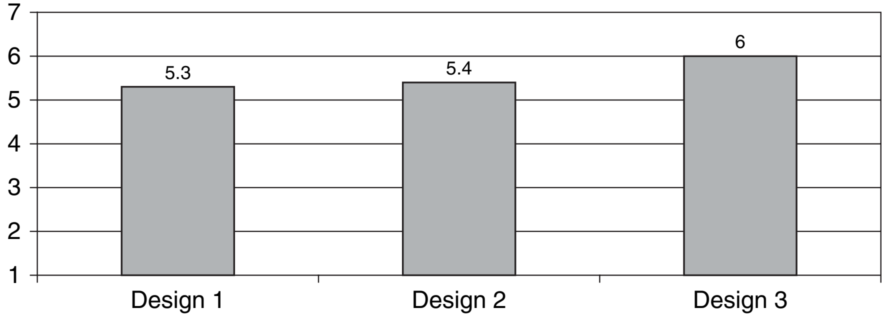


FIGURE 10.6

Phase 2: Credibility scores of three different website designs.

didn't seem to mind advertising as long as it was related to the site's content. The conditions that scored lower in attractiveness and had the highest levels of unrelated ads scored lowest in credibility.

Negative Reactions

I measured the number of unprompted negative comments made by users during the study. Results showed that Design 1 (many unrelated ads) prompted the most negative reactions (see Figures 10.7 and 10.8). These findings suggested that although the number of ads could affect participant experience, the relevancy of the ads had a stronger effect on people's perception of the site. Ads that were

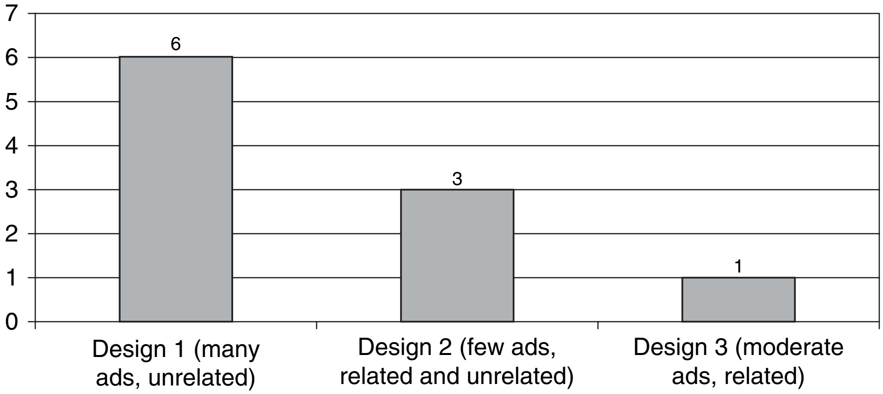


FIGURE 10.7

Number of negative comments. Design 1 received the most negative comments. Having many unrelated ads on a website generated the most negative reactions.

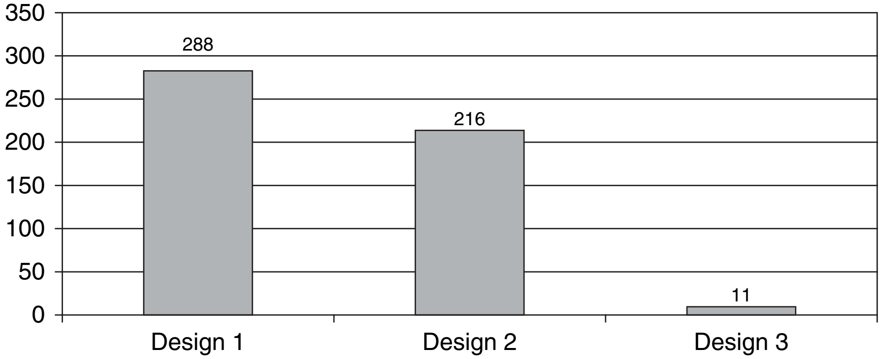


FIGURE 10.8

Seconds spent on negative comments. Respondents spent the least amount of time making negative comments when using Design 3, whereas Designs 1 and 2 evoked longer negative reactions.

unrelated to the site's content evoked stronger and more negative reactions than designs that had related advertising content.

Outcome: Phase 2

We were not particularly interested in statistical significance; we were interested in trends and design insight. Even with feedback from only seven participants, the findings from this study strongly influenced the team's decision on which visual design to keep and which ones to scrap. This study also helped influence advertising guidelines, such as the placement, design, and types of advertisement allowed on the site.

10.1.3 Phase 3: Testing a Single Design

By Phase 3, the team felt relatively confident that they had a solid design. The designers made minor modifications to the design that respondents preferred and put it through another round of usability testing. This phase had even fewer participants—only five—but this was fine because the process allowed for several validation points. Similar tasks were given to the new set of respondents. Again, success scores and user interactions were noted. At the end of each session, participants were asked to rate their satisfaction level with using the website.

Results: Phase 3

This design resulted in an average satisfaction rating of 5.8 out of 7. Overall, people were relatively successful in accomplishing the tasks. The average success score across all tasks was 92 percent (see Figure 10.9).

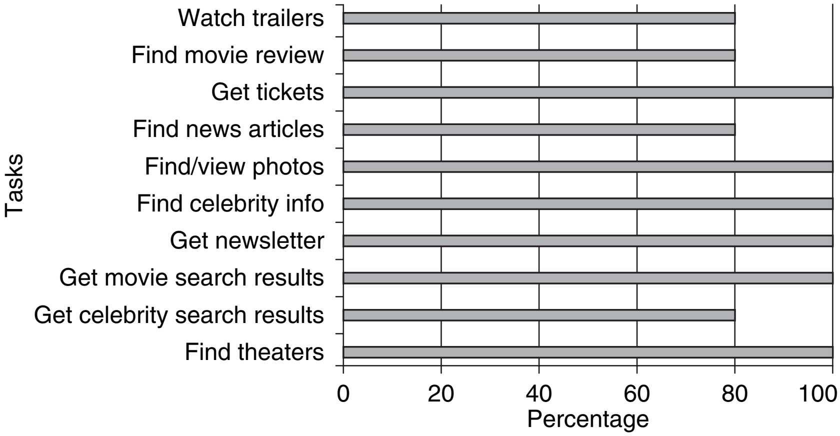


FIGURE 10.9

Average success rate. The high rate of success gave us a boost in confidence—we were heading in the right direction.

Outcome: Phase 3

The positive results from this study were an indication that we were on the right path. The data suggested that people can easily accomplish tasks on the site. Our diligence paid off. The satisfaction score was slightly lower than I originally expected, but this could be due to the lack of functionality available on the prototype. For example, people were disappointed when the search engine was not available, and this might have affected their rating.

10.1.4 Conclusion

In a few short months, the team was able to completely redesign an entertainment content site to meet both user and business goals. Augmenting quantitative techniques with simple quantitative methodologies provided sufficient data to help the design team make decisions quickly. Only seven days total were spent user testing, and only 27 people participated in the study. However, even with such seemingly small numbers, the data were sufficient to guide us in making the right decisions. After Phase 3, the design went through another round of iterations. I conducted a design review, and the new website was launched shortly after with great success.

10.1.5 Biography

Hoa Loranger is a User Experience Specialist at Nielsen Norman Group, where she consults with many large, well-known companies in various industries, including finance, entertainment, technology, e-commerce, and mobile devices. She co-authored the book *Prioritizing Web Usability* (New Riders Press). Her extensive research has spanned the globe, including Asia, Australia, and Europe. She is a frequent speaker and has published reports on a variety of web usability topics.

10.2 USABILITY EVALUATION OF A SPEECH RECOGNITION IVR—By James R. Lewis

The purpose of this study was to investigate the immediate usability of a speech recognition interactive voice response (IVR) system. The application supports self-service and call-routing activities, such as finding nearby stores, finding nearby service centers, and purchasing products and accessories.

10.2.1 Method

Six participants completed the following four tasks.

- *Task 1*: Find the nearest service location
- *Task 2*: Find an accessory and a nearby store
- *Task 3*: Buy an accessory
- *Task 4*: Order a new product

After each task, participants completed an After Scenario Questionnaire (ASQ; Lewis, 1995). After finishing all tasks, participants completed a Post-Study System Usability Questionnaire (PSSUQ; Lewis, 1995, 2002).

The ASQ contains three questions, each scored from 1 to 7, with lower scores indicating greater satisfaction and ease of use. The overall score is the arithmetic mean of the three item scores. The PSSUQ contains 16 items for which participants indicate their level of agreement on a 7-point scale, with lower ratings indicating greater subjective usability. The PSSUQ has three subscales: System Usefulness (SysUse), Information Quality (InfoQual), and Interface Quality (IntQual).

Each participant read through each of the tasks, asked for any necessary clarifications, and indicated to the experimenter when he or she was ready to begin the tasks. Participants received instruction to complete an ASQ following each task.

10.2.2 Results: Task-Level Measurements

For each task, the mean call-completion time was less than three minutes and the upper limits of the confidence intervals (CIs) were all under three and a half minutes (see Table 10.1). It's difficult to interpret time-on-task without a comparative context, but the observed times-on-task can provide a baseline for future studies of similar systems.

Participants completed Tasks 2, 3, and 4 with 100 percent success. The failures for Task 1 occurred when two participants incorrectly selected "Find a store" instead of "Find service" from the main menu.

The satisfaction scores (7-point scales in which 1 is the best rating) for Tasks 1, 2, and 3 were very good. The rating for Task 4 was slightly poorer, primarily due to confusion over the wording of a main menu option. Five of the six participants indicated some feeling of confusion about the option, even though they all eventually chose the correct option. In addition to this, two participants spoke subsets of the prompted phrase that were out of grammar and not recognized by the application.

Table 10.1 Task-Level Usability Measurements

Task	Completion Time		Success Rate		Satisfaction	
	Mean	90% CI	Rate	90% CI	Mean	90% CI
1	2.1	1.8–2.3	0.67	0.27–0.94	1.5	1.0–2.0
2	2.9	2.3–3.4	1.00	0.61–1.00	1.9	1.2–2.6
3	2.6	2.0–3.2	1.00	0.61–1.00	2.1	1.0–3.1
4	1.3	0.9–1.6	1.00	0.61–1.00	2.4	0.9–3.9

Table 10.2 PSSUQ Data by Scale

Satisfaction	Mean	90% CI
Overall	2.0	1.0–3.1
SysUse	2.0	1.0–3.1
InfoQual	2.2	1.0–3.3
IntQual	1.8	1.0–2.8

10.2.3 PSSUQ

The overall mean PSSUQ score (with 90 percent CI) was 2.0 ±1.1. Table 10.2 gives the means and 90 percent confidence intervals for the PSSUQ scales. The observed means were well below the scale midpoint (in the direction of favorable perceived usability), and the upper limits of their 90 percent confidence intervals were also below the scale midpoint, indicating that for this design, these tasks, and these types of users, the perception of usability was favorable.

10.2.4 Participant Comments

Each participant provided comments for the three most-liked and least-liked system attributes, as shown in Table 10.3. The most frequently mentioned application characteristics that participants liked were that the application was easy to understand, easy to use, had good recognition accuracy, and had a useful repeat function. The most frequently mentioned dislikes were that one of the main menu options was confusing, some information (especially telephone numbers) was spoken too quickly, and some of the information presented was too wordy.

Table 10.3 Participant Comments

Participant Comments: Favorable	Count
Easy to understand (1, 2, 5, 6)	4
Easy to use (3, 4, 6)	3
Good recognition (1, 3)	2
Able to have system repeat (2, 4)	2
Able to barge in (4)	1
Quick (1)	1
Prerecorded audio (2)	1

Table 10.3 *cont...*

Participant Comments: Unfavorable	Count
One main menu prompt terminology confusing (1, 2, 4)	3
Some information was spoken too quickly (1, 5)	2
Too wordy (1, 5)	2
Found barge-in didn't always work (6)	1
Lack of DTMF input (6)	1
Voice recognition unforgiving (5)	1

Note: Participants' numbers are in parentheses.

10.2.5 Usability Problems

Table 10.4 lists the usability problems discovered during this study, organized by task. The numbers in the cells of the participant columns indicate the impact level of the problem on that participant, using the following scale:

1. Caused participant to fail to complete the task successfully
2. Required more than one minute to recover from the problem
3. Required less than one minute to recover from the problem
4. Minor inefficiency

Table 10.5 shows the usability problems, organized in order of descending severity. The severity score is calculated by multiplying the frequency of occurrence of the problem in the study (the percentage of participants who experienced the problem) by an impact weight. Greater impact levels receive greater weight, with Level 1 having a weight of 10, Level 2 having a weight of 5, Level 3 having a weight of 2, and Level 4 having a weight of 1. Thus, the highest possible severity score is 1,000 (frequency of 100 times an impact weight of 10).

10.2.6 Adequacy of Sample Size

This assessment of the adequacy of the sample size of a problem-discovery (formative) usability study uses the relatively new techniques described in Lewis (2006). An analysis of the participant by problem matrix in Table 10.4 indicated that the observed rate of problem discovery was 0.278 and, adjusting for the sample size and pattern of problem discovery, produced an adjusted estimate of 0.134. Given $p = 0.134$, the best estimate of the percentage of discovered problems is 0.578

Table 10.4 Usability Problems by Participant

Problem Number	Description	Task	P1	P2	P3	P4	P5	P6
1	System did not recognize zip code.	1		2				
2	Asked for store rather than service location.	1				1	1	
3	Phone numbers play a little fast.	1	4	4			4	
4	Had to repeat phone number.	2		3				
5	Picked store from list, didn't wait for prompt to request zip code.	2			4			
6	Experienced "Was that <silence>?"	2				3		
7	System did not recognize an ungrammatical utterance.	2						3
8	Tried to enter zip with keypad.	2						3
9	Went back to main menu to order accessory.	3		3				
10	Participant said entire model number, then rejected 4-digit "Was that" response.	3					3	
11	A main menu option was confusing.	4	3	3	3	3		3
12	Participant said ungrammatical variant of prompted option, had to listen to help and retry.	4					3	3

$[1 - (1 - 0.134)^6]$. At this rate of discovery, the best estimate of the total number of problems available for discovery given this application and the tasks is about 21 (12 / 0.578). Table 10.6 shows the expected rate of continuing problem discovery.

This projection shows that the 12 problems discovered in this study by these six participants represent about 57.8 percent of the problems available for discovery. At this relatively slow rate of discovery, it is likely that running additional participants will only turn up one additional problem per participant for the next two or three participants and will then continue declining until it becomes necessary to run multiple participants to discover each new problem. This low rate of problem discovery and generally low problem severities are indicative of a well-designed user interface in which there are few high-frequency problems to find.

Table 10.5 Usability Problems by Severity

Problem Number	Description	Task	Impact Level	Frequency	Impact Weight	Severity
2	Asked for store rather than service location.	1	1	33	10	333
11	A main menu option was confusing.	4	3	83	2	167
1	System did not recognize zip code.	1	2	17	5	83
12	Participant said ungrammatical variant of prompted option, had to listen to help and retry.	4	3	33	2	67
3	Phone numbers play a little fast.	1	4		1	50
4	Had to repeat phone number.	2	3	17	2	33
6	Experienced “Was that <silence>?"	2	3		2	33
7	System did not recognize an ungrammatical utterance.	2	3	17	2	33
8	Tried to enter zip with keypad.	2	3	17	2	33
9	Went back to main menu to order accessory.	3	3	17	2	33
10	Participant said entire model number, then rejected 4-digit “Was that” response.	3	3	17	2	33
5	Picked store from list, didn’t wait for prompt to request zip code.	2	4	17	1	17

Table 10.6 Expected Rate of Continuing Problem Discovery

N	*p*	Problems	Delta
7	0.635	13.3	1.2
8	0.684	14.4	1.0
9	0.726	15.2	0.9
10	0.763	16.0	0.8
11	0.795	16.7	0.7
12	0.822	17.3	0.6
13	0.846	17.8	0.5
14	0.867	18.2	0.4
15	0.884	18.6	0.4
16	0.900	18.9	0.3
17	0.913	19.2	0.3
18	0.925	19.4	0.2
19	0.935	19.6	0.2
20	0.944	19.8	0.2
21	0.951	20.0	0.2

10.2.7 Recommendations Based on Participant Behaviors and Comments

Recommendation 1: Investigate reordering the main menu options to reduce the likelihood that a user seeking service will select ‘‘Find a store’’ when the desired function is ‘‘Find service.’’ The basis for this recommendation is that two of the six participants made this high-impact error when completing Task 1. Placing ‘‘Find service’’ immediately after ‘‘Find a store’’ will increase the likelihood that callers will hear both options before saying, ‘‘Find a store.’’ (We have observed that when callers barge in, they often do so after hearing one option past the option that they ultimately select.)

Recommendation 2: Replace the confusing main menu option with alternative wording. Five of six participants indicated that they found one of the main menu options somewhat confusing.

Recommendation 3: Slightly slow the text-to-speech (TTS) speech rate, especially for telephone numbers. Two participants commented that some of the speech was too fast, and three participants specifically commented on the speed of playback of telephone numbers, which is done via TTS. One participant had the application repeat a phone number, presumably due to the speed of the speech.

Recommendation 4: Consider allowing zip code entry with the keypad. One participant tried to enter a zip code using the keypad rather than speech. If the impact to code is slight, then enable the keypad for zip code entry. If the impact to code is high, then leave the application as is because the participant recovered rapidly and completed the task successfully.

10.2.8 Discussion

This case study of a formative usability study illustrates a number of different usability metrics. Even though the focus was on problem discovery, it didn't require much additional investment to provide the standard task-level measurements of completion time, success rate, and satisfaction rating (and the study-level satisfaction ratings of the PSSUQ). To document the limits of measurement accuracy with a sample size of six participants, the report included 90 percent confidence intervals for completion time and satisfaction ratings and 90 percent binomial confidence intervals for success rates. From a usability perspective, the task- and study-level measurements are of limited immediate usefulness but could be useful as benchmarks for tests of later versions of the application or tests of similar tasks in other similar applications.

The primary focus of the study was on discovering and prioritizing usability problems and developing recommendations to eliminate the problems or reduce their impact. At the usability problem level of analysis, the key measurements were frequency of occurrence and impact on the user, with those measurements combined to derive a severity score used to prioritize the usability problems.

Finally, I assessed the adequacy of the sample size for the purpose of problem discovery. The techniques used for this assessment are fairly new, but I now routinely apply them at the end of these types of studies to get an idea of whether it is reasonable to run additional participants in the study. For the development group with whom I currently work, the presentation of results in tabular form is very effective, so I only rarely present results in graphic form.

10.2.9 Biography

James R. (Jim) Lewis has worked as a Human Factors Engineer for IBM since 1981. He has published research on sample sizes for formative usability studies and the measurement of user satisfaction. He served in 2005 as the chair for the NIST working group on formative usability metrics, and wrote the chapter on usability testing for the 3rd edition (2006) of the *Handbook of Human Factors and Ergonomics*.

10.2.10 References

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*, 57–78.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*, 463–488.

Lewis, J. R. (2006). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1275–1316). New York: John Wiley.

10.3 REDESIGN OF THE CDC.gov WEBSITE—By Robert Bailey, Cari Wolfson, and Janice Nall

In 2006, the Centers for Disease Control and Prevention (CDC) elected to revise and update its website, CDC.gov. One of the primary goals during this redesign process was to optimize the usability of the homepage and a few new second-level pages.

A 6-person User Experience team—Janice Nall, Robert Bailey, Catherine Jamal, Colleen Jones, Nick Sabadosh, and Cari Wolfson—was organized and headed by Janice Nall. The team created a plan to ensure that major, meaningful usability-related activities would be appropriately applied. Over a 6-month period, the major usability activities included the following:

- Conducting a review of past usability studies at CDC
- Interviewing users, stakeholders, partners, and web staff
- Conducting a detailed analysis of web and search and call logs
- Analyzing the user survey data from the American Customer Satisfaction Index
- Surveying the ideas and attitudes of CDC leadership, employees, and web staff
- Conducting a card-sort activity
- Conducting parallel design sessions
- Producing a series of wireframes
- Creating graphically oriented prototypes

This case study focuses only on the major usability activity of usability testing, particularly as it was related to revising the CDC.gov homepage. The usability testing included a baseline test, first-click tests, and final prototype tests. Overall, 170 participants were tested using more than 100 task scenarios in three major usability tests. These usability tests eventually showed a success rate improvement of 26 percent and a satisfaction score improvement of 70 percent.

10.3.1 Usability Testing Levels

We used the model of usability testing levels Bailey developed in 2006 to help guide our decisions about the types of usability tests to perform. This model proposes five usability testing levels:

Level 1: Traditional inspection evaluations, such as heuristic evaluations, expert reviews, and so forth

Level 2: Algorithmic reviews with scenarios

Level 3: Usability tests that are moderately controlled and use a relatively small number of test participants (about eight)

Level 4: Usability tests that are tightly controlled but use only enough participants to make weak inferences to the population

Level 5: Usability tests that are very tightly controlled and use a sufficient number of participants to make strong inferences to the population

Because of the well-documented weaknesses of inspection evaluations (Bailey, 2004), the User Experience team elected not to use any Level 1 testing and to do Level 2 testing only on the final, revised homepage. The final algorithmic evaluation was based on the usability guidelines book (Koyani, Bailey, & Nall, 2006).

The existing CDC.gov homepage had been around with very few changes since February 2003 (almost four years). During this time, there had been many surveys, studies, and tests, recommending small changes to the homepage. We were interested in collecting data where we could make fairly strong inferences to the user population, and consequently we did few Level 3 tests. Most of our usability tests were either Level 4 or Level 5.

10.3.2 Baseline Test

The original baseline test was used to establish the human performance and user satisfaction levels for the existing site. We also used the baseline test to help us understand some of the major usability issues that may help to guide future changes to the new homepage. The baseline tests took place in August 2006 at three different locations in the United States. The in-person usability tests were conducted in government usability labs, conference rooms, and offices.

The participants were both federal employees and people who had no affiliation with the government. We tested a total of 68 participants that included public health professionals; healthcare workers (physicians, nurses, etc.); general consumers; researchers and scientists; and journalists, legislators, and students. The participants had a mix of gender, age, education, race, and Internet experience that matched typical users of CDC.gov. The usability testing sessions were about one hour long and were conducted using Keynote's *WebEffective* and Techsmith's *Morae*.

The User Experience team created 36 scenarios that reflected the tasks performed most frequently on CDC.gov. Considerable time and effort went into identifying the most frequently performed tasks on the website. Information was used from interviews, surveys, reports from the call center, evaluations of web logs (Omniture), ACSI results, and so on.

Each participant dealt with ten scenarios. All participants were told to browse to the correct answer, and not to use the website's search capability. Later, a ''search-only'' test was conducted to determine the impact (if any) of not allowing users to search. As can be seen in Table 10.7, there was little difference in success rates, average time, average page views, or satisfaction scores.

Participants were instructed to work as quickly and accurately as possible. They answered prescenario questions, responded to several task scenarios, and then answered postscenario questions (including a satisfaction metric: the System Usability Scale—SUS). If they did not find the answer to a scenario question within three minutes, they were automatically moved to the next scenario and were considered ''unsuccessful'' on that scenario.

The overall success rate across all 36 scenarios was 54 percent. In other words, users were able to successfully complete only about half of the scenarios in the allowed 3-minute time limit. Figure 10.10 shows success rates for 11 of the scenarios.

10.3.3 Task Scenarios

Using information gained during the baseline test, we focused on making changes to those scenarios that elicited the poorest performance. For those

Table 10.7 Overall Performance Comparisons between "Browse" and "Search" Activities

	Browse	Search
Success rate	54%	49%
Average time	2.4 min	2.8 min
Average page views	7.1 pages	7.7 pages
Satisfaction (out of 100)	46	49

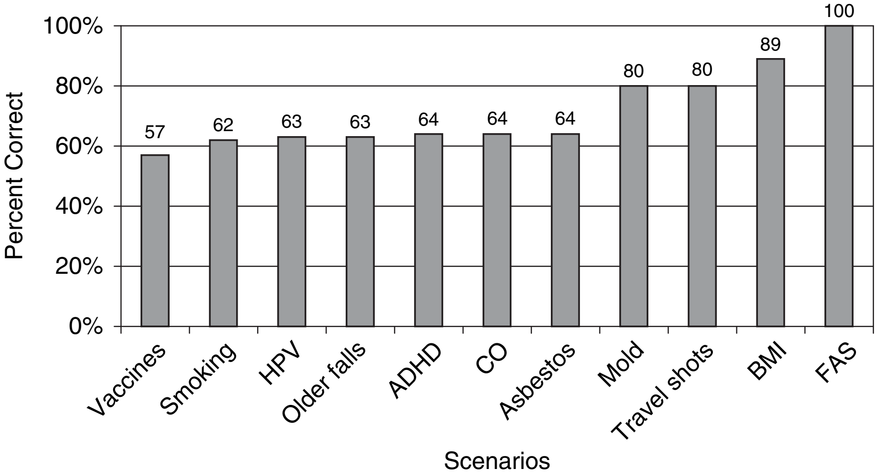


FIGURE 10.10

Success rates of eleven scenarios from the baseline test.

scenarios shown in Figure 10.10, this would be those that had success rates of 64 percent or less. As can be seen from Figure 10.10, there were some scenarios that elicited very good performance. We tried to encourage designers *not* to make changes that would lower the success rates of these scenarios. We found it necessary to continuously remind designers to stay focused on those scenarios where the success rate was lowest, and to look for general reasons why they were low. This was one of the major uses of the usability testing sessions: It kept designers focused on those homepage issues that most needed their attention.

10.3.4 Qualitative Findings

It should be noted that the qualitative findings from the usability testing were used to inform the proposed changes to the homepage. We had participants type their overall impressions, specify what they liked best and least, and indicate what changes they would make if given the chance. During many of the tests, evaluators took notes about problems that participants were having and asked questions in a debriefing at the end of each test. Here are some of the observations after the baseline test:

- Many felt that the homepage had too much information (overwhelming).
- Participants struggled to find information because of busy, cluttered pages.
- Participants who found the A–Z index liked it and used it quite frequently (it was hard to find).

- Participants thought that the website was inconsistent in layout, navigation, and look and feel.
- Participants did not feel that the categories of information were clear.
- Participants thought that they had to go through too many layers to find information.
- Participants *did* find that the features and the page descriptions were useful.

We found these observations to be invaluable after deciding which scenarios were leading to the most usability problems. These insights from both testers and users assisted us in deciding what changes had the best chance of improving the website.

10.3.5 Wireframing and FirstClick Testing

Once the problems were better understood and several solutions had been proposed, we created several competing wireframes to see which would best elicit the success levels we were seeking. All of these eventually were combined into three homepage wireframes (A, B, and C). After only one quick test, B was eliminated and the final two wireframes (A and C) were tested "head-to-head" (see Figure 10.11).

We had 65 participants attempt to complete 136 scenarios (68 using Wireframe A and 68 using Wireframe C). Each participant spent about one hour completing the scenarios using Wireframe A and then Wireframe C (or Wireframe C and then Wireframe A).

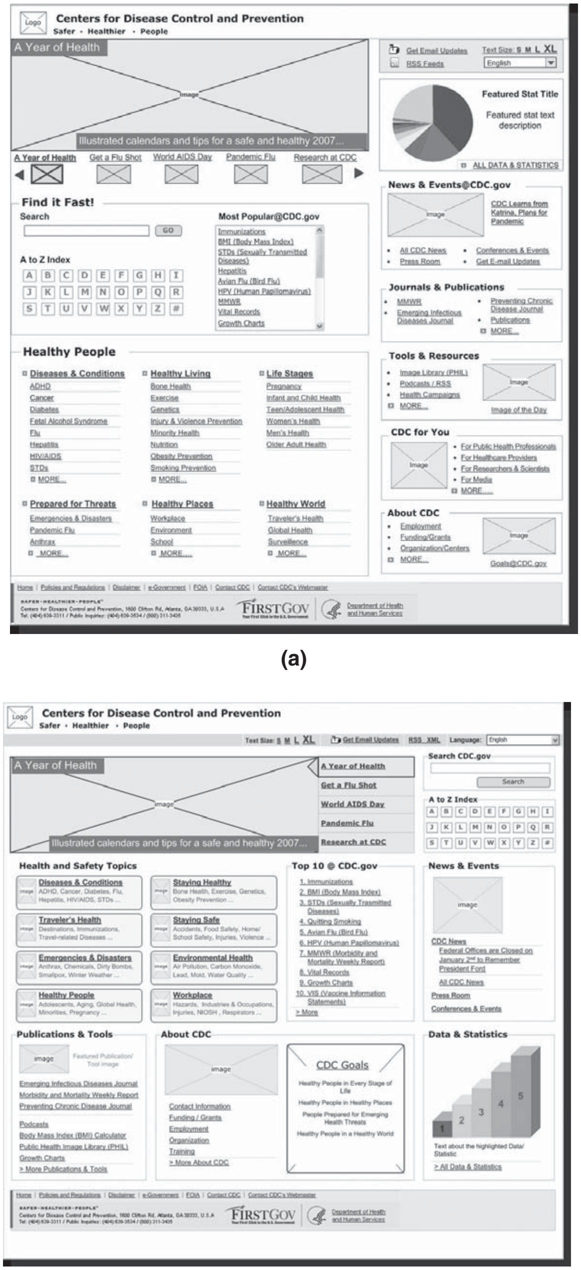
We used a FirstClick testing methodology, where we collected and analyzed only the first click they made after reading the scenario. In previous testing, we had observed that the first click was a very critical action. If they had difficulty with that original decision, they frequently had problems finding the correct answer, and when the first click was wrong, it took much longer to complete a scenario than it should have.

This type of testing enabled us to considerably expand the number of scenarios (each scenario took participants less than 30 seconds) and to see which of the two wireframes elicited the best *initial* performance. The test was conducted using Bailey's *Usability Testing Environment* and Techsmith's *Morae*.

The test results showed no reliable difference between the two wireframes in terms of success, but Wireframe C did elicit reliably *faster* performance. Figures 10.12 and 10.13 show the percent of participants that clicked first on each of the links for two different scenarios. Those with a light background were correct, and those with a dark background were incorrect.

We were particularly interested in two different response patterns. Figure 10.12 shows a scenario where a fairly large number of participants tended to agree on the *wrong* response—in other words, the wrong link.

This means that there is something about that link that erroneously elicits clicks, whereas the correct response does not elicit the clicks. This usually can be fixed by


Logo
Centers for Disease Control and Prevention
Safer · Healthier · People
A Year of Health
image
Illustrated calendars and tips for a safe and healthy 2007...
A Year of Health
Get a Flu Shot
World AIDS Day
Pandemic Flu
Research at CDC
Find it Fast!
Search
GO
A to Z Index
A B C D E F G H I
J K L M N O P Q R
S T U V W X Y Z #
Most Popular@CDC.gov
Immunizations
BMI (Body Mass Index)
STDs (Sexually Transmitted Diseases)
Hepatitis
Avian Flu (Bird Flu)
HPV (Human Papillomavirus)
MMWR
Vital Records
Growth Charts
Healthy People
Diseases & Conditions
ADHD
Cancer
Diabetes
Fetal Alcohol Syndrome
Flu
Hepatitis
HIV/AIDS
STDs
MORE...
Healthy Living
Bone Health
Exercise
Genetics
Injury & Violence Prevention
Minority Health
Nutrition
Obesity Prevention
Smoking Prevention
MORE...
Life Stages
Pregnancy
Infant and Child Health
Teen/Adolescent Health
Women's Health
Men's Health
Older Adult Health
Prepared for Threats
Emergencies & Disasters
Pandemic Flu
Anthrax
MORE...
Healthy Places
Workplace
Environment
School
MORE...
Healthy World
Traveler's Health
Global Health
Surveillence
MORE...
Get Email Updates
RSS Feeds
Text Size: S M L XL
English
Featured Stat Title
Featured stat text description
ALL DATA & STATISTICS
News & Events@CDC.gov
Image
CDC Learns from Katrina, Plans for Pandemic
All CDC News
Press Room
Conferences & Events
Get E-mail Updates
Journals & Publications
MMWR
Emerging Infectious Diseases Journal
Preventing Chronic Disease Journal
Publications
MORE...
Tools & Resources
Image Library (PHIL)
Podcasts / RSS
Health Campaigns
MORE...
Image
Image of the Day
CDC for You
Image
For Public Health Professionals
For Healthcare Providers
For Researchers & Scientists
For Media
MORE.....
About CDC
Employment
Funding/Grants
Organization/Centers
MORE...
Image
Goals@CDC.gov
Home | Policies and Regulations | Disclaimer | e-Government | FOIA | Contact CDC | Contact CDC's Webmaster
SAFER·HEALTHIER·PEOPLE™
Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333, U.S.A
Tel: (404) 639-3311 / Public Inquiries: (404) 639-3534 / (800) 311-3435
FIRSTGOV
Your First Click to the U.S. Government
Department of Health and Human Services
Logo
Centers for Disease Control and Prevention
Safer · Healthier · People
Text Size: S M L XL
Get Email Updates
RSS XML
Language:
English
A Year of Health
image
Illustrated calendars and tips for a safe and healthy 2007...
A Year of Health
Get a Flu Shot
World AIDS Day
Pandemic Flu
Research at CDC
Search CDC.gov
Search
A to Z Index
A B C D E F G H I
J K L M N O P Q R
S T U V W X Y Z #
Health and Safety Topics
Diseases & Conditions
ADHD, Cancer, Diabetes, Flu, Hepatitis, HIV/AIDS, STDs ...
Staying Healthy
Bone Health, Exercise, Genetics, Obesity Prevention ...
Traveler's Health
Destinations, Immunizations, Travel-related Diseases ...
Staying Safe
Accidents, Food Safety, Home/ School Safety, Injuries, Violence ...
Emergencies & Disasters
Anthrax, Chemicals, Dirty Bombs, Smallpox, Winter Weather ...
Environmental Health
Air Pollution, Carbon Monoxide, Lead, Mold, Water Quality ...
Healthy People
Adolescents, Aging, Global Health, Minorities, Pregnancy ...
Workplace
Hazards, Industries & Occupations, Injuries, NIOSH , Respirators ...
image
Top 10 @ CDC.gov
1. Immunizations
2. BMI (Body Mass Index)
3. STDs (Sexually Trasmitted Diseases)
4. Quitting Smoking
5. Avian Flu (Bird Flu)
6. HPV (Human Papillomavirus)
7. MMWR (Morbidity and Mortality Weekly Report)
8. Vital Records
9. Growth Charts
10. VIS (Vaccine Information Statements)
> More
News & Events
image
CDC News
Federal Offices are Closed on January 2nd to Remember President Ford
All CDC News
Press Room
Conferences & Events
Publications & Tools
image
Featured Publication/ Tool image
Emerging Infectious Diseases Journal
Morbidity and Mortality Weekly Report
Preventing Chronic Disease Journal
Podcasts
Body Mass Index (BMI) Calculator
Public Health Image Library (PHIL)
Growth Charts
> More Publications & Tools
About CDC
image
Contact Information
Funding / Grants
Employment
Organization
Training
> More About CDC
CDC Goals
Healthy People in Every Stage of Life
Healthy People in Healthy Places
People Prepared for Emerging Health Threats
Healthy People in a Healthy World
Data & Statistics
1
2
3
4
5
Text about the highlighted Data/ Statistic
> All Data & Statistics
Home | Policies and Regulations | Disclaimer | e-Government | FOIA | Contact CDC | Contact CDC's Webmaster
SAFER·HEALTHIER·PEOPLE™
Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333, U.S.A
Tel: (404) 639-3311 / Public Inquiries: (404) 639-3534 / (800) 311-3435
FIRSTGOV
Department of Health and Human Services

(a)

(b)

FIGURE 10.11

The two wireframes used for conducting the FirstClick tests.

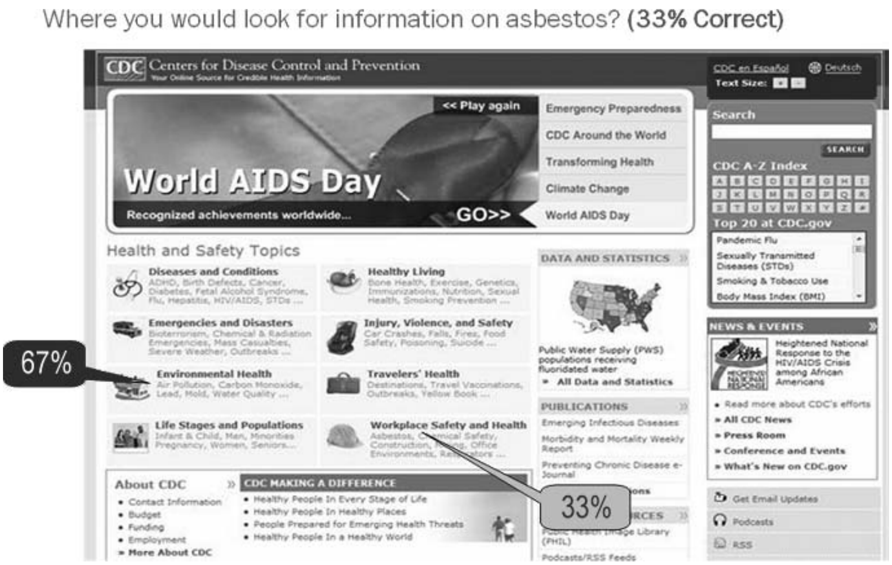

Where you would look for information on asbestos? **(33% Correct)**
67%
33%
CDC Centers for Disease Control and Prevention
Your Online Source for Credible Health Information
CDC en Español
Deutsch
Text Size:
<< Play again
World AIDS Day
Recognized achievements worldwide...
GO>>
Emergency Preparedness
CDC Around the World
Transforming Health
Climate Change
World AIDS Day
Search
SEARCH
CDC A-Z Index
Top 20 at CDC.gov
Pandemic Flu
Sexually Transmitted Diseases (STDs)
Smoking & Tobacco Use
Body Mass Index (BMI)
Health and Safety Topics
Diseases and Conditions
ADHD, Birth Defects, Cancer, Diabetes, Fetal Alcohol Syndrome, Flu, Hepatitis, HIV/AIDS, STDs ...
Healthy Living
Bone Health, Exercise, Genetics, Immunizations, Nutrition, Sexual Health, Smoking Prevention ...
Emergencies and Disasters
Bioterrorism, Chemical & Radiation Emergencies, Mass Casualties, Severe Weather, Outbreaks ...
Injury, Violence, and Safety
Car Crashes, Falls, Fires, Food Safety, Poisoning, Suicide ...
Environmental Health
Air Pollution, Carbon Monoxide, Lead, Mold, Water Quality ...
Travelers' Health
Destinations, Travel Vaccinations, Outbreaks, Yellow Book ...
Life Stages and Populations
Infant & Child, Men, Minorities Pregnancy, Women, Seniors...
Workplace Safety and Health
Asbestos, Chemical Safety, Construction, Office Environments, Respirators ...
DATA AND STATISTICS
Public Water Supply (PWS) populations receiving fluoridated water
All Data and Statistics
PUBLICATIONS
Emerging Infectious Diseases
Morbidity and Mortality Weekly Report
Preventing Chronic Disease e-Journal
RESOURCES
Public Health Image Library (PHIL)
Podcasts/RSS Feeds
NEWS & EVENTS
Heightened National Response to the HIV/AIDS Crisis among African Americans
Read more about CDC's efforts
All CDC News
Press Room
Conference and Events
What's New on CDC.gov
Get Email Updates
Podcasts
RSS
About CDC
Contact Information
Budget
Funding
Employment
More About CDC
CDC MAKING A DIFFERENCE
Healthy People In Every Stage of Life
Healthy People In Healthy Places
People Prepared for Emerging Health Threats
Healthy People In a Healthy World

FIGURE 10.12

Sixty-seven percent of the participants had an incorrect FirstClick response (call-out percent with dark background).

changing the link names. Figure 10.13 shows a different pattern of responses. In this case, few people could agree on the correct response, and even fewer on the incorrect response. Those making wrong responses showed little consistency in their responses. It is much more difficult to find a workable solution for a problem like this.

Some of the scenarios that elicited poor success rates in the baseline test continued to show poor performance in the FirstClick test. This was after making many changes to the homepage that were expected to improve user performance on these scenarios. For example, in the baseline test "Budget" had an overall success rate of only 17 percent, and in the FirstClick test "Budget" still had only 59 percent making the first click successfully. We noted that two scenarios that elicited good success rates in the baseline test continued to show very good performance in the FirstClick test. The FAS scenario, for example, had perfect performance in both tests. For some scenarios there were big differences between the performance levels on the two different prototypes. These were evaluated further.

10.3.6 Final Prototype Testing (Prelaunch Test)

Some design decisions on both wireframes led to better performance, and we used the best from both wireframes to produce a final wireframe. The final wireframe

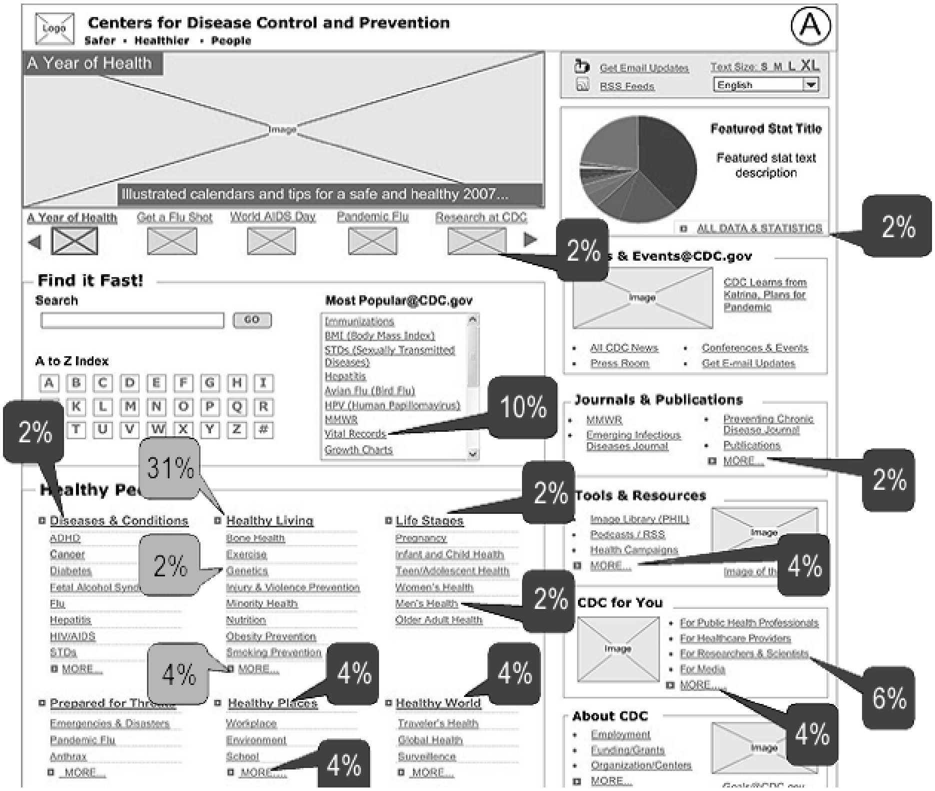


FIGURE 10.13

Most participants were incorrect in their FirstClick response. (Call outs with a light background were correct; call outs with a dark background were incorrect.)

was then used as the basis for developing a graphic prototype. The graphic prototype, with images, colors, and a variety of type fonts for headers, was used for the final prelaunch usability test.

The participants were a mixture of federal and nonfederal employees. The majority of participants included the primary audience for this site, which were healthcare providers, public health professionals, and consumers. Again, these tests were conducted using Bailey's *Usability Testing Environment*.

The prelaunch usability tests were divided into two parts: a pilot test and the final test. The pilot test was conducted on Monday using 18 participants and 56 task scenarios that were divided into three categories: 24 FirstClick from the homepage, 24 FirstClick from one of the new second-level pages, and 9 "homepage to content page." All participants saw all scenarios in one-hour in-person testing sessions. After the testing was complete, the data were summarized and analyzed, and a set of observations and recommendations were prepared. On

Tuesday morning, the usability team met with the primary designers and conveyed the recommendations. On Tuesday afternoon and evening, many changes were made to the homepage, and a few changes were made to the test itself. The final in-person tests were conducted all day on Wednesday, using 56 slightly revised scenarios and 19 different participants.

For purposes of this case study, we will only discuss the findings related to the nine ‘‘homepage to content page’’ scenarios in the final in-person (prelaunch) tests. In the ‘‘homepage to content page’’ scenarios, users were allowed to navigate through the website to find information, exactly as they had done in the original baseline test. This provided us with an estimate of the percent improvement from the baseline to the final prelaunch test. The results showed a success rate of 78 percent that could be compared back to the original success rate of 62 percent for these same scenarios (an improvement of 26 percent). In addition, the satisfaction score was measured as 78, which could be compared with the original satisfaction score of 46 (an improvement of 70 percent). A summary of improvements is shown in Table 10.8. Notice that all human performance and user satisfaction scores were substantially improved.

As can be seen in Figure 10.14, the changes made to the homepage for some of these nine scenarios led to substantial increases in performance (Budget, ADHD, and BMI). Four others had moderate increases in performance. Eight of the nine scenarios (89 percent) showed improved performance. Looking back at the FirstClick data, two of these scenarios were shown to elicit good performance (BMI and FAS), whereas one of them still showed that it was a problem (Budget). Obviously, useful changes were made to the website to bring the success rate for ‘‘Budget’’ from 17 percent in the baseline test to 59 percent in the FirstClick test to 91 percent in the prelaunch test.

Only one of these scenarios had a decrease in performance: ‘‘NewsRelease.’’ The ‘‘NewsRelease’’ scenario shows how certain changes can result in no substantial changes to the success rate, whereas others actually can cause difficulties. The success rate for ‘‘NewsRelease’’ went from a low 67 percent in the baseline to an

Table 10.8 Comparison of Performance and Preference Results between Baseline and Prelaunch Tests

	Baseline Test	Prelaunch Test	Percent Improvement
Success rate	62	78	26%
Average time	96	81	19%
Average page views	8.3	4.9	69%
Satisfaction score (out of 100)	46	78	70%

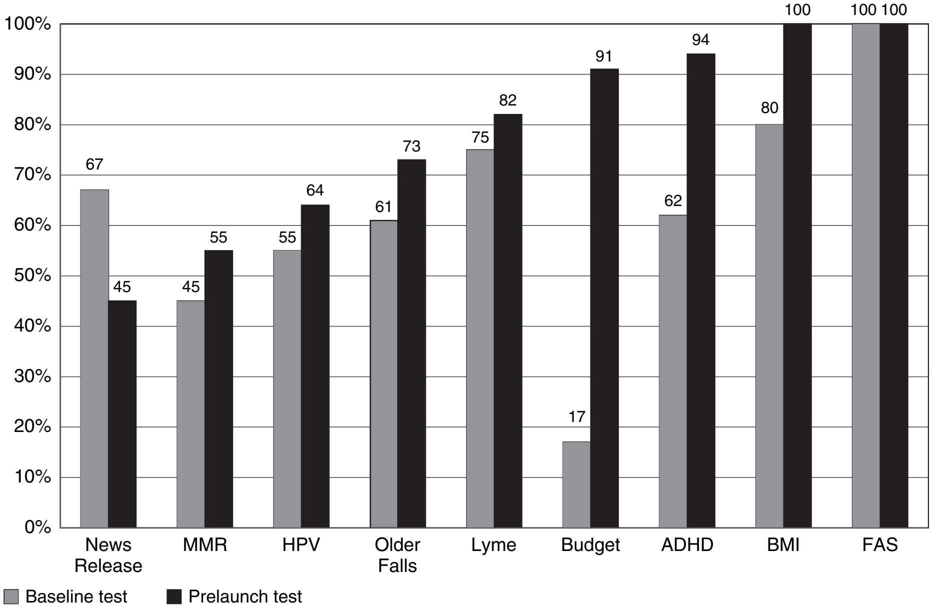


FIGURE 10.14

Change in performance between the baseline test and the prelaunch test.

even lower 48 percent in the FirstClick to a 45 percent in the prelaunch test. None of these success rates were acceptable, and this task continues to pose a problem for users who are looking for news information on the homepage.

10.3.7 Conclusions

There are three main uses of usability testing metrics in the revision of existing web pages and websites. The first is to establish an overall baseline success rate for comparison against other sites, usability objectives, a revised website, and so on.

The second major use is to identify the individual scenarios where users have the most problems. This helps in finding the weakest areas of the site and enables designers to focus the majority of their efforts on strengthening those areas. In other words, a good usability test can help guide designers to the most problematic areas of the site. We continuously stayed focused on those scenarios that had the lowest success rates. This also enabled us to look for patterns across several scenarios. For example, making the A–Z Index more apparent had a positive impact on several scenarios. In another situation, we noted that virtually all of the CDC statistics-related scenarios performed poorly.

With each change to the website and the resulting test results, we watched to see if success increased. Once the success rate reached the goal for each scenario, we stopped trying to improve performance on that scenario. We tried to use our limited resources and to stay focused on those scenarios that were causing the most usability problems.

A third use of the usability test metrics was to identify those scenarios that consistently elicited high levels of human performance. These areas do *not* require attention, and focusing on them would only waste time and detract from more useful activities.

After each test, changes were made to the homepage. With each set of usability tests and resulting changes made to the homepage (i.e., each "iteration"), we watched closely to see if the success level of the targeted scenarios increased, decreased, or remained about the same. We knew from past testing that some changes to the website would lead to improved performance, others would have no effect, and others could actually make things worse. This is the major reason why it is so important *always* to test after making changes to the site.

In the CDC.gov revision of the homepage, and certain second-level pages, we found the usability testing program to be very valuable. The new homepage was substantially improved over the original page. Based on our experiences, we find it hard to imagine how valid, useful, and important changes could be made to improve a website without conducting high-level usability testing to assist the change process.

10.3.8 Biographies

Robert W. Bailey is the president of Computer Psychology, Inc. He holds a Ph.D. degree from Rice University, worked for years in the Human Performance Technology Center at Bell Laboratories, and has served on the faculties of numerous schools, including Columbia University. Dr. Bailey has an international reputation as an author, researcher, and lecturer.

Cari A. Wolfson is president of Focus on U!, a usability and user-centered design firm that has led the redesign of several high-profile websites, including Usability.gov. Prior to founding Focus on U!, Wolfson was a Usability Engineer at the National Cancer Institute and a Senior Information Architect with a large Internet consulting firm. Wolfson holds a B.A. in communications from Drake University.

Janice R. Nall is the director of eHealth Marketing at the National Center for Health Marketing at the Centers for Disease Control and Prevention (CDC). She is responsible for CDC's websites and leading e-health and new media initiatives. Previously, she managed user experience activities at the General Services Administration (GSA), and National Cancer Institute (NCI).

10.3.9 References

Bailey, R. W. (2006). Applying usability metrics. *Web Managers University,* May 16.

Bailey, R. W. (2004). Comparing heuristic evaluation and performance testing. *User Interface Update*.

Koyani, S. J., Bailey, R. W., and Nall, J. R. (2006). *Research-based web design & usability guidelines*. Washington, DC: U.S. Government Printing Office.

10.4 USABILITY BENCHMARKING: MOBILE MUSIC AND VIDEO—By Scott Weiss and Chris Whitby

Usable Products Company studied U.S. mobile phone-based download and playback of video and music by way of a comparative, quantitative user experience benchmark. The studied mobile phone user interfaces for music and video playback were poor, despite the availability of high-quality dedicated media players in the marketplace. Despite polished visual design, usability participants had a difficult time purchasing and playing video and music on mobile phones.

10.4.1 Project Goals and Methods

This research focused on understanding the strengths and weaknesses of today's implemented means for shopping, purchase, and playback of video and music on mobile handsets. In the first and second quarters of 2006 in New York City, we compared the three media download and playback services available for mobile handsets by way of usability interviews with target users of these services. At the same time, we utilized improvements in our benchmarking process and learned from new mistakes made in information presentation. For the first time, we included video clips in the presentation files, but the sheer density and quantity of information prevented our participants from finding what they needed quickly.

10.4.2 Qualitative and Quantitative Data

This study combined quantitative and qualitative methods to compare three mobile handsets from different carriers and manufacturers. The quantitative data provided the core of the evidence for usability findings (Ebling & John, 2000). The quantitative data and statistical analysis provided support for and scope of the qualitative issues (Sauer & Kindlund, 2005). The quantitative metrics provided points of comparison for each of the three handsets tested and uncovered issues not discovered through qualitative means.

10.4.3 Research Domain

A limited number of handsets were available that supported both music and video downloads at the time of this study. Only three operators, one of which was an MVNO (Mobile Virtual Network Operator), supported the technology. All were CDMA, as high-speed GSM networks were not yet launched in the United States. Verizon and Sprint featured multiple handsets, and we chose the most comparable handsets from two major manufacturers, Samsung and LG. Amp'd Mobile had only one handset at the time, from Kyocera, which was available in two colors. We chose the Angel, a pearlescent white slider phone.

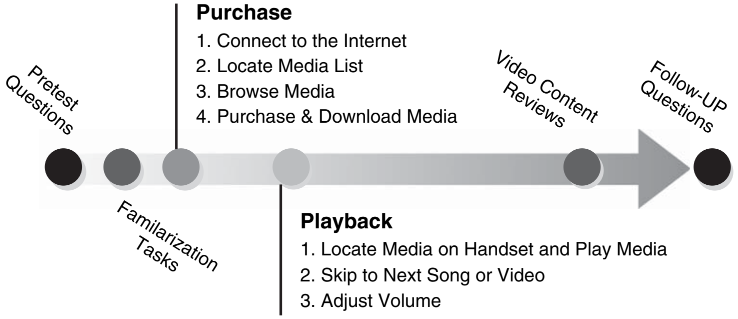


FIGURE 10.15

Purchase and playback tasks used in the study.

10.4.4 Comparative Analysis

Respondent tasks were categorized as either "purchase" or "playback" (see Figure 10.15). By collecting time, success rate, and satisfaction data for each task, not only were we able to get a more granular level of quantitative data that pertained to particular task flows, but we were also able to compare tasks flows between handsets and between different tasks on the same handset.

Connect to the Internet was presented to respondents as "Find a place where you can start to browse for clips that you can purchase." The *Media List* was "a list of media, or media categories." These two segments were distinct in the Amp'd Kyocera Angel but were combined on the Sprint and Verizon handsets, which immediately brought up Media Lists.

Respondents were first asked to find and play back the song, clip, or channel they had purchased. Then they were asked to play the *next* piece of content of the same type (song, video) and adjust the playback volume. The playback task was repeated on the two clamshell handsets that featured external media controls to determine the usability differential between internal and external user interfaces (UIs). Each respondent performed two sets of tasks, once with music and once with video. The music and video tasks were alternated between respondents.

10.4.5 Study Operations: Number of Respondents

Research published on sample sizes and user testing focuses on using online methods to extend lab-based research. Spool and Schroeder (2001) discussed how the size and complexity of websites require additional research participants to "ferret out problems." Schulman (2001) argued that quantitative research provides the answers to this problem through online usability testing methods. Online testing is valid for desktop PC software and websites, but it cannot help in the case of testing multiple mobile telephone handsets. In order to acquire the data for mobile telephones, which cannot be instrumented like websites or PC software, in-person one-on-one usability interviews are required. For this project, an ‘‘interview’’ was a user testing session during which performance and preference measurements were taken. Participants were asked to complete tasks under observation in a one-on-one laboratory setting.

Our study was comparative in nature, so it was important to prevent ‘‘single-user effects’’ from coloring these averages. Based on past experience and successful case studies similar in nature, we arrived at 20 as an acceptable compromise for the number of user interviews per handset.

10.4.6 Respondent Recruiting

Sixty respondent interviews were required to complete the research (three handsets at 20 interviews per handset). To meet this recruiting goal, we developed an online panel early in the pilot phase and our panel development was continuous throughout the research.

We targeted mobile phone users representing a mix of gender, age, ethnicity, household income, educational level, and mobile phone data use, skewed slightly toward younger people, who more actively purchase handset-based media. Each panelist completed an extensive online survey designed to facilitate the recruiting process. A combination of database applications, e-mail, and telephone was used in the recruiting effort.

We developed customized recruiting software for this project (see Figure 10.21), which interfaced the respondent database with the project’s data warehouse. This custom UI was used by an administrative assistant to filter, schedule, and track respondents for the study. The software generated a web calendar of respondents for moderators who were in separate buildings in New York City.

10.4.7 Data Collection

We used the freely available logging tool from Ovo Studios (*http://www.ovo.com*) during interviews. This tool enabled systematic data capture, covering performance and perception measures, summarized in Table 10.9. Free-form moderator observations that formed the basis for qualitative analysis were also captured with this logging tool, which also provided task times and observation categories. Synching the software’s timer with that of the DVD recorders capturing the interview allowed the moderator to set checkpoints in the logs after a potentially presentable video clip (a segment during which a respondent properly articulates the perceived general sentiment about an aspect of a handset’s usability) so they could be harvested afterwards. Any ambiguities or missing data in the logs could be retrieved by reviewing the DVD recordings of interviews.

Table 10.9 Performance and Perception Metrics Collected

Performance Metrics	Perception Metrics
■ Time to complete tasks ■ Success rates ■ Number of attempts	■ Feelings about handsets before and after one hour's use (affinity) ■ Perceived cost ■ Perceived weight of handset ■ Ease of use ■ Perceived time to complete tasks ■ Satisfaction

10.4.8 Time to Complete

Successful tasks were timed from a predefined start point to a predefined stop point. For example, the Download Place segment started from when a respondent started to use the handset after receiving instructions and lasted until the respondent successfully connected to the media service. At two minutes, respondents were offered help. If they accepted help, the task was failed. Without help, tasks were failed at three minutes.

10.4.9 Success or Failure

Success was determined when a respondent achieved the requested goal within the time limit.

10.4.10 Number of Attempts

In our previous studies, a respondent's attempt count was incremented by one each time he or she returned to the handset's idle screen while trying to complete a task. However, after a case study conducted by Usable Products Company between 2004 and 2005, we stopped recording attempt counts because successful respondents were almost always completing a task on their first or second attempt. Respondents who exceeded two attempts either gave up on the task or ran out of allotted time (Martin & Weiss, 2006). For the purpose of this study, the moderator did not maintain an attempt count, but rather he carefully monitored and recorded the actions of respondents as they tried to complete tasks.

10.4.11 Perception Metrics

As shown in Table 10.9, preferences were also measured, all on a 5-point scale, with 1 being the worst measure and 5 being the best measure. We found the ‘‘Before and After Affinity'' to be most interesting, since the change in the measure from before the interview to after was affected only by participant experience of the handset. We asked respondents to state how they felt about the handset, from "1 = Hate It" to "5 = Love It."

10.4.12 Qualitative Findings

Qualitative analysis followed the typical usability reporting format, with issues and illustrations arranged by task. With this arrangement, useful comparisons could be made between the particular media paths on a handset, and comparisons for the same task stage could be made across handsets.

To provide extra value, the report included Information Architecture diagrams with photographs of the display mapping the task flows for all handsets. These diagrams were actual screenshots from the handsets arranged with linking arrows to walk readers through typical usage patterns for task completion.

10.4.13 Quantitative Findings

Presentation of the qualitative data included charts for every category of quantitative data collected. In all, we produced 73 charts, perhaps too many for our readership. Our readers informed us that they would prefer a more condensed analysis of our findings. We utilized the Summative Usability Metric (SUM) method for single-scoring and ranking usability (see Figure 10.16). In future studies we will produce fewer charts and rely more heavily on SUM scores.

10.4.14 Summary Findings and SUM Metrics

In studying usability publications, we came upon the Summative Usability Metric, based on the work of Sauro and Kindlund (2005). In this method, time, success, and ease of use perception data are normalized with respect to benchmark goals and averaged, giving a value on a Six Sigma scale. Based on the assumption of a normal distribution, SUM charts show how likely the task performance is to exceed the ideal measures we set for these tasks.

In Figure 10.16, the Sprint LG Fusic clearly had the best usability with regard to adjusting the volume. The worst usability came from the Amp'd Kyocera Angel, with regard to locating and playing downloaded songs. Analysis of the data in the benchmark required that we look to the information architecture diagrams and individual handset qualitative analyses. In each outlier case, there were specific design flaws in the user experience that explained the handset's poor performance in the SUM chart, validating the measure from a logical standpoint.

10.4.15 Data Manipulation and Visualization

Data was stored using the Microsoft Access database and analyzed with SPSS, often considered the industry standard in statistical software. Microsoft Access

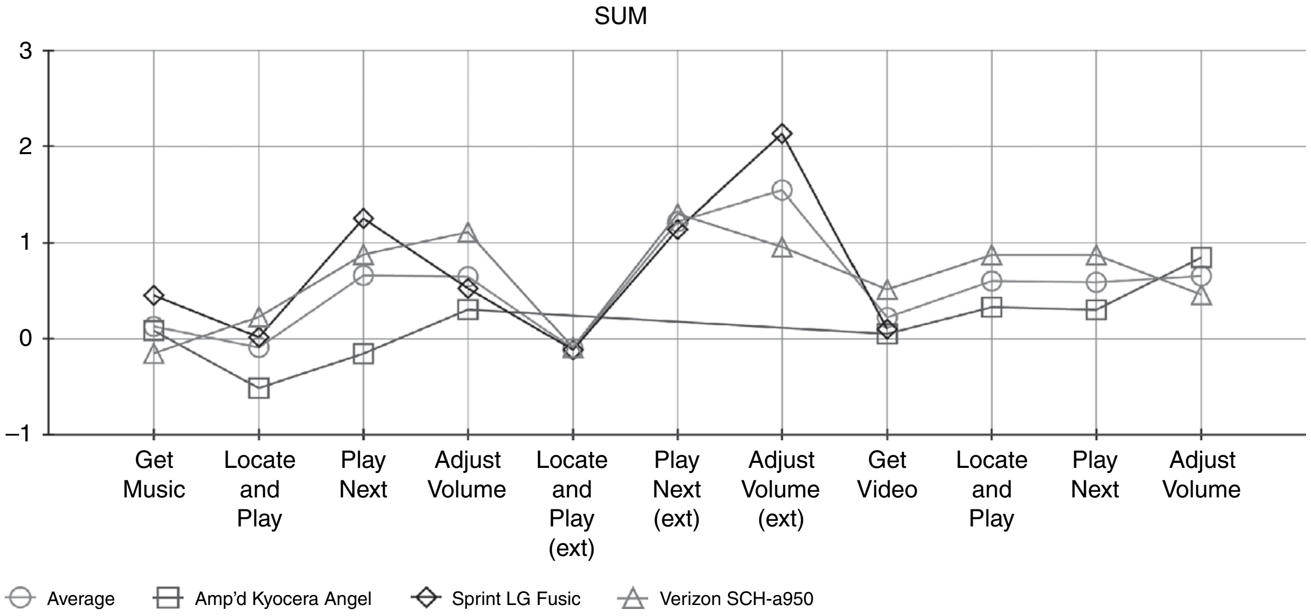


FIGURE 10.16

SUM metrics for the three handsets studied.

compartmentalized the data storage and processes stages in the analysis and sped up the data integrity checks and quality assurance work done before report delivery. Using SPSS rather than Microsoft Excel allows one to increase the number of statistical procedures available without adding to development time, outside basic scripting.

Charts were designed and coded with Visual Basic for Applications (VBA) in the Microsoft Access environment, relying on the Scalable Vector Graphics (SVG) format for output. The SVG markup language provides precise control of vector graphics via an easily scripted markup language. Since SVG is a full graphic platform and not a charting plug-in for Microsoft Access, crafting the visualizations and charting conventions became a much simpler process, and we were able to incorporate the visualizations into our reports in a high-quality fashion.

Video clips of respondents making significant remarks were harvested from the DVD recordings of interviews before being edited for quality using Adobe Premier.

10.4.16 Discussion

The end result of the study was a set of best practices that explained common problems respondents encountered while attempting music and video download and playback tasks. The documentation included embedded respondent video clips and flowchart diagrams using screenshots to help visualize the usability challenges of the handsets.

The main usability challenge was the vague link between the devices' hard buttons and their corresponding functions. For example, the functionality of the Amp'd Kyocera Angel's "Back" button was inconsistent between user interfaces. When the Amp'd Media Player was accessed on the device, the "Back" button's functionality became "broken." During playback of a media file, the "Back" button only served to move playback to the start of the file, not to return to the previous screen, which confused respondents, who spent more time navigating than would otherwise have been necessary.

It was unclear to respondents that the download place might not be accessible through the web browser, as was the case with the Verizon and Sprint handsets. Respondents, who associated downloading media with the interim step of "connecting," were confused when the Verizon and Sprint web browsers did not provide access to media services.

While browsing, differentiating between types of media was difficult for respondents when using the Amp'd Live service. Music was easily identified, but the "TV" category included TV-show themed wallpaper in addition to video clips.

Respondents also had difficulty locating downloaded media on all three handsets. The option to immediately view or listen to the media immediately after download at the purchase confirmation screen was unavailable, forcing respondents to then search for the media before playback. It was nonobvious to respondents that downloaded songs and video could be found through the purchasing service; instead, they checked local media folders (e.g., "My Content") and media players before checking the purchase place as a last resort. The Amp'd Kyocera Angel employed cryptic filenames (e.g., "UMG_0034561"), further confusing respondents.

During playback, respondents were unable to switch between local media files without exiting the media player. Although all tested players had the ability to switch songs within the player, the barrier we discovered was the default playlists that were used when the players were opened. Players that opened songs without including other downloaded songs on the playlist had, in effect, disabled their navigation buttons, and their single-song playlists were useless for switching between songs.

10.4.17 Benchmark Changes and Future Work

Our clients indicated their preference for more concise findings. We erred on the side of providing comprehensive analysis to prove the points made in our best practices. In future work, we will use summaries and best practices for the early matter and provide more detailed evidence for findings later in the documentation. Some charts will simply be left out if they are of minimal value. We will also use larger type throughout the document, requiring fewer words and more illustrations. Larger type benefits everyone—print and screen readers alike.

We found that the data collected in this project would have been nice to compare to our earlier syndicated study of ring tone, wallpaper, and game downloads and installation. However, the data from that project was siloed, preventing easy cross-project analysis. We will embark on a data warehousing project to consolidate usability projects from multiple projects so that we can expand our analysis capabilities. For our next research, we plan to study either mobile search or mobile community user interfaces.

10.4.18 Biographies

Scott Weiss is the president of Usable Products Company, a usability design and research agency specializing in usability benchmarking and UI design for mobile phones and consumer electronics. Usable Products worked with Samsung on the design of the YP-K5 MP3 player UI, and with Vodafone on the Simply handset UI. Scott's prior employers were Apple, Microsoft, Sybase, and Autodesk.

Chris Whitby is a Usability Analyst with Usable Products Company. His responsibilities include discussion guide development, respondent recruitment, moderating, data analysis, and presenting findings. Chris graduated from New York University with a bachelor of arts degree in sociology.

10.4.19 References

Ebling, M. R., and John, B.E. (2000). On the contributions of different empirical data in usability testing. *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques,* 289–296.

Martin, R., and Weiss, S. (2006). Usability benchmarking case study: Media downloads via mobile phones in the U.S. *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services,* 195–198.

Sauro, J. (2004). Premium usability: Getting the discount without paying the price. *Interactions*, *11*(4), 30–37.

Sauro, J., and Kindlund, E. (2005). A method to standardize usability metrics into a single score. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 401–409.

Schulman, D. (2001). *Quantitative usability: Extending lab research for larger sample sizes*. Retrieved January 14, 2005, from *http://www.3.ibm.com/ibm/easy/eou_ext.nsf/Publish/1854/$File/1854.pdf* .

Spool, J., and Schroeder, W. (2001). Testing web sites: Five users is nowhere near enough. *CHI 2001 Proceedings,* 286–286.

10.5 MEASURING THE EFFECTS OF DRUG LABEL DESIGN AND SIMILARITY ON PHARMACISTS' PERFORMANCE—By Agnieszka (Aga) Bojko

In an effort to make drug packaging production more cost effective, Abbott created a standard template for its prescription drug labels. Although replacing the myriad labels for Abbott's drugs appeared daunting, the initiative also provided the opportunity to add new anticounterfeiting features to the label. The standardization would also help reduce regulatory review time of new drugs.

Throughout the redesign process, Abbott focused on safety because any label change could easily add to the estimated 1.7 percent of dispensing errors that occur in pharmacies (Flynn, Barker, & Carnahan, 2003). To ensure that the template did not have a negative impact on pharmacists' performance relative to the existing labels, User Centric was asked to conduct a series of studies that investigated various applications and elements of the new label.

Initial studies showed that the new label, when applied to drug bottles, improved performance in terms of both search efficiency and information-processing efficiency. These improvements led to shorter times required to locate and identify critical drug information with the new labels as compared to the existing labels (Bojko, Gaddy, Lew, Quinn, & Israelski, 2005). So what exactly caused the improvement? Applying a template to all Abbott drug labels meant two things: change in label design (the visual design of the template was different from all existing label designs) and increase in consistency across labels (all labels were created according to the template). Although we knew that the template had a positive effect, we did not know exactly why. Was the observed improvement due to the new design or the interlabel similarity?

The main objective of the present study was to investigate whether the proposed standardized label template had an impact on user performance when the template was applied to drug cartons (containing tablets or capsules individually sealed in blister packs) rather than drug bottles (containing loose tablets or capsules). If there was an impact, we also wanted to determine whether it was due to the different label design, the increased similarity across all Abbott labels, or some combination of these two factors.

Since bottles are much more common than cartons in U.S. pharmacies, we decided to conduct the study in Europe, where cartons are the primary form of prescription drug packaging. Pharmacists were invited to our user research lab in Rome and asked to perform typical drug selection tasks using the new and existing label designs in a situation that approximated their work environment.

Considering Abbott's concern for safety, the ideal metric to assess user performance would be error rate. However, drug-dispensing errors attributed to label design are quite rare and presumably would be even less frequent in a lab setting. Thus, although we needed to collect error data, we also decided to measure factors that can *contribute* to error, such as difficulties in finding information or increased cognitive processing demands. The effect of these factors can be accentuated and lead to errors in stressful situations and under high mental workload.

Therefore, in addition to error rate, we measured the number of eye fixations prior to target selection as an indicator of search efficiency (Kotval & Goldberg, 1998), average fixation duration as a measure of information processing difficulty (Fitts, Jones, & Milton, 1950), and pupil diameter of the participants as a global measure of cognitive workload (Kahneman, 1973). As a gross measure of overall task efficiency, we also analyzed participant response times needed to locate and select the correct drug.

10.5.1 Participants

Twenty pharmacists (14 women and 6 men) between the ages of 26 and 45 ($M = 34$) participated in individual 60-minute sessions. Their pharmaceutical experience ranged from 2 to 17 years ($M = 6.5$), and they worked in pharmacies of varying sizes (3–13 employees) and script volumes (150−1,500 scripts per day). The participants were compensated for their time with a one-year subscription to a professional journal.

10.5.2 Apparatus

The stimuli were presented on a 17-inch monitor interfaced with a PC with a 1.79 GHz AMD Athlon XP 2100+ processor. The screen resolution was set to 1024 × 768 pixels. Each participant used a keyboard to indicate responses. Eye movements were recorded with a Tobii 1750 binocular remote eye-tracker with 50 Hz temporal resolution and a .5-degree spatial resolution.

10.5.3 Stimuli

In countries where cartons are more prevalent than bottles, prescription drugs tend to be stored in drawers, as shown in Figure 10.17. In an effort to simulate the

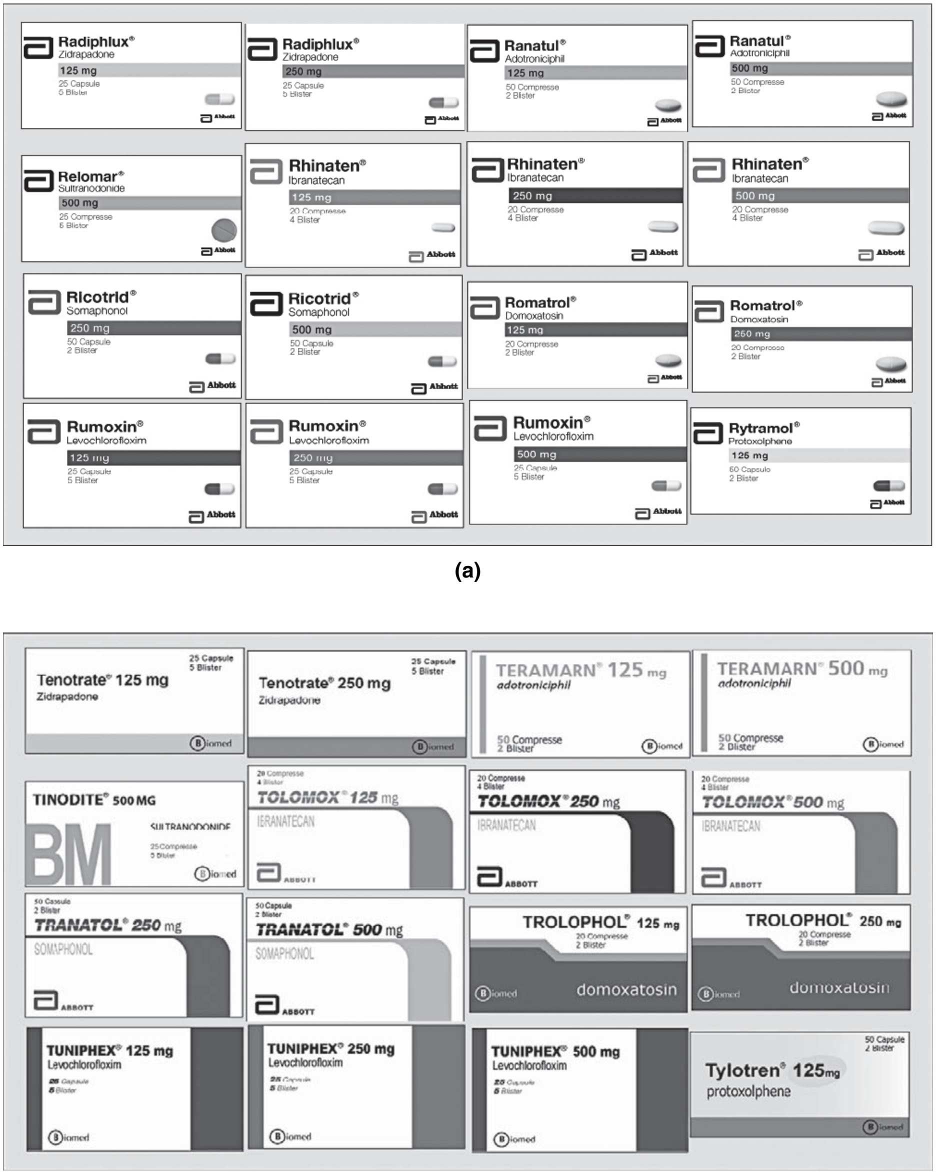
FIGURE 10.17

Drawers with cartons in an Italian pharmacy.

experience of standing over a drawer and selecting a drug to fill a prescription, we created 16 digital drawings of carton labels arranged in a 4 × 4 grid representing a top-down view of a drawer, examples of which are shown in Figure 10.18. Each drawer stimulus contained 16 cartons with 8 unique drugs, some including multiple dosage strengths. The arrangement of the labels corresponded to a common drug organization—alphabetically by trade name, followed by dosage strength within the name.

Four different Abbott label designs were tested: the new template and three existing designs (E1, E2, and E3), with two label similarity levels (low and high), resulting in the eight experimental conditions illustrated in Figure 10.19. Each condition consisted of a set of two drawers. The high-similarity condition shown in Figure 10.18(a) was achieved by using the same label design as a template for all labels in the drawer. In the low-similarity condition shown in Figure 10.18(b), the Abbott label design of interest was used for only five cartons per drawer (for two different drugs). The other 11 cartons (six different drugs) used various non-Abbott labels, presented as products of a fictitious manufacturer ‘‘Biomed.’’

The eight sets of drawers (one set per condition) were equivalent in terms of color and all label elements. All elements except for the drugs’ trade names were mapped exactly (e.g., in one of the drawers in each set, the top left label was always yellow with 125 mg, as shown in Figure 10.19). The trade names were different in each drawer, but they all consisted of exactly three syllables with a name length of seven to ten letters. All trade and generic drug names were fictitious to control for familiarity effects and ensure content equivalency across drawers by allowing us to systematically manipulate the different label elements without confusing the participants.


Radiphlux® Zidrapadone 125 mg 25 Capsule 5 Blister
Radiphlux® Zidrapadone 250 mg 25 Capsule 5 Blister
Ranatul® Adotroniciphil 125 mg 50 Compresse 2 Blister
Ranatul® Adotroniciphil 500 mg 50 Compresse 2 Blister
Relomar® Sultranodonide 500 mg 25 Compresse 5 Blister
Rhinaten® Ibranatecan 125 mg 20 Compresse 4 Blister
Rhinaten® Ibranatecan 250 mg 20 Compresse 4 Blister
Rhinaten® Ibranatecan 500 mg 20 Compresse 4 Blister
Ricotrid® Somaphonol 250 mg 50 Capsule 2 Blister
Ricotrid® Somaphonol 500 mg 50 Capsule 2 Blister
Romatrol® Domoxatosin 125 mg 20 Compresse 2 Blister
Romatrol® Domoxatosin 250 mg 20 Compresse 2 Blister
Rumoxin® Levochloroflоxim 125 mg 25 Capsule 5 Blister
Rumoxin® Levochlorofloxim 250 mg 25 Capsule 5 Blister
Rumoxin® Levochlorofloxim 500 mg 25 Capsule 5 Blister
Rytramol® Protoxolphene 125 mg 60 Capsule 2 Blister
Abbott
(a)
Tenotrate® 125 mg Zidrapadone 25 Capsule 5 Blister
Tenotrate® 250 mg Zidrapadone 25 Capsule 5 Blister
TERAMARN® 125 mg adotroniciphil 50 Compresse 2 Blister
TERAMARN® 500 mg adotroniciphil 50 Compresse 2 Blister
TINODITE® 500 MG SULTRANODONIDE 25 Compresse 5 Blister BM
TOLOMOX® 125 mg IBRANATECAN 20 Compresse 4 Blister ABBOTT
TOLOMOX® 250 mg IBRANATECAN 20 Compresse 4 Blister ABBOTT
TOLOMOX® 500 mg IBRANATECAN 20 Compresse 4 Blister ABBOTT
TRANATOL® 250 mg SOMAPHONOL 50 Capsule 2 Blister ABBOTT
TRANATOL® 500 mg SOMAPHONOL 50 Capsule 2 Blister ABBOTT
TROLOPHOL® 125 mg 20 Compresse 2 Blister domoxatosin
TROLOPHOL® 250 mg 20 Compresse 2 Blister domoxatosin
TUNIPHEX® 125 mg Levochlorofloxim 25 Capsule 5 Blister
TUNIPHEX® 250 mg Levochlorofloxim 25 Capsule 5 Blister
TUNIPHEX® 500 mg Levochlorofloxim 25 Capsule 5 Blister
Tylotren® 125mg protoxolphene 50 Capsule 2 Blister
Biomed

(a)

(b)

FIGURE 10.18

Two of the 16 stimuli used in the study: (a) high-similarity drawer with the new labels; (b) low-similarity drawer with one of the existing label designs—E1.

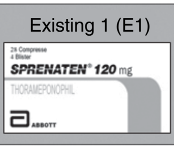
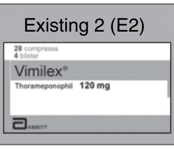
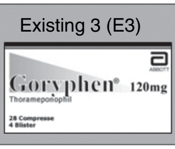
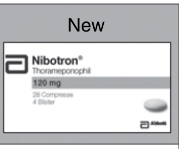
		Design			
		Existing 1 (E1) 	Existing 2 (E2) 	Existing 3 (E3) 	New 
Similarity	**High**	Example not shown	Example not shown	Example not shown	**Example in Figure 10.18(a)**
	Low	**Example in Figure 10.18(b)**	Example not shown	Example not shown	Example not shown

FIGURE 10.19

The eight experimental conditions.

Only three different dosage strengths were presented in a single drawer (either 125, 250, and 500 mg, or 120, 240, and 480 mg). This forced users to examine both the drug name and dosage strength because the target could not be identified by looking at one of these elements alone.

10.5.4 Procedure

For each task, participants saw a picture of a "drawer" and were asked to locate a drug according to task instructions (e.g., "Find Medotil 500 mg"). The instructions were always displayed at the bottom of the screen. After reading the instructions, participants had to press the space bar for the drawer stimulus to appear above the instructions. As soon as the participants identified the target label, they responded by pressing one of four keys on the keyboard (*d*, *f*, *k*, and *l*) to indicate the column (I, II, III, or IV) in which the target drug was located. The key press began the next trial by displaying a new set of instructions.

Participants used their index and middle fingers from each hand to press the keys, which were marked with stickers to make them easy to locate. We used key press responses instead of mouse clicks to ensure that our visual search data were uncontaminated by the selection behavior associated with aiming the mouse pointer.

Each condition was presented in a separate block, which consisted of 16 trials (tasks), with 8 trials per drawer. Each trial had a unique target. In each low-similarity drawer, 4 trials had an Abbott label as a target and 4 trials had a non-Abbott label as a target. In each high-similarity drawer, all 8 trials had an Abbott label as a target because all labels used the same Abbott template.

To compensate for order effects, the order of the trials within each block was randomized, and the order of the blocks was counterbalanced across participants. Organization of the drugs in each drawer remained constant, as is usually the case in pharmacies. Participants were instructed to find the target drug as quickly as possible but without sacrificing accuracy. To familiarize them with the procedure and response mappings, a practice block of 16 trials was administered at the beginning of the session.

10.5.5 Analysis

The data we collected included error rate, time-on-task, and three eye-related measures: pupil diameter, fixation count, and average fixation duration. We precisely defined the metrics to ensure valid results:

Error rate was defined as the percentage of responses (key presses) that did not match the location of the target in question.

Time-on-task was the time it took participants to read task instructions, locate the correct label, and respond. It was measured from the onset of the instructions to the participant's key press indicating the location of the drug in question, and it was computed for successful tasks only.

Pupil diameter was the average size of the participant's pupil throughout the entire block (condition). Pupils dilate slowly, so collecting this measure on a per-trial basis (or excluding data from unsuccessful trials) would not be useful because of the difficulty in matching the data to the particular stimulus caused by the delay.

Fixation count was the average number of fixations per Abbott label. Fixations were determined using the dispersion-threshold method (Salvucci & Goldberg, 2000), where the dispersion threshold radius was 0.5° and duration threshold was 100 ms. This measure was computed for successful tasks only.

Fixation duration was the average length of a fixation on Abbott labels. This measure was computed for all tasks (both successful and not).

An initial analysis of the fixation data revealed that each Abbott label in the low-similarity condition received, on average, more fixations than each of the Abbott labels in the high-similarity condition. However, on closer examination of the data, we realized that this was an artifact caused by the much higher target-to-Abbott-label ratio in the low-similarity condition (4 targets/5 Abbott labels) than in the high-similarity condition (8 targets/16 Abbott labels).

Targets tended to attract more attention because participants had to examine both the name and dosage strength of the drug to make sure that the label matched the drug in the instructions (while elimination of nontargets was possible by looking at one of these elements only). Also, participants often employed strategies where once they found the correct drug, they would look at the instructions again, and then go back to the label to verify. These two factors significantly increased the number of fixations on the labels that served as targets, thus increasing the average for the Abbott labels in the low-similarity condition as compared to the high-similarity condition.

To avoid this artifact in the fixation analysis, we decided to analyze only those labels in the high-similarity drawers that matched the locations of the five Abbott labels in the low-similarity drawers. The chosen locations, in addition to being equivalent along multiple dimensions (as described in Section 10.5.3), also had the same probability of being a target (i.e., four of five of these labels were always targets on the eight trials with each drawer).

Once all measures were defined, to determine the impact of the new label template on pharmacists' performance, we computed a 4 (design: E1, E2, E3, and New) × 2 (similarity level: low and high) repeated measures analysis of variance (ANOVA) for all five metrics collected during the study.

10.5.6 Results and Discussion

On average, participants needed 9 seconds to read the task instructions and correctly identify the target drug among 15 others. Each participant responded erroneously on 5.9 percent of the tasks, which translates into almost one error per condition. However, at least some of the errors resulted from an incorrect response rather than an incorrect intention because participants would often comment that they pressed the wrong key unintentionally. The error and time analyses revealed no main effects for design or label similarity and no significant interaction between the two factors ($p > 0.05$), which suggests that the new template had no negative impact at least on the overt aspects of participant performance.

Participant pupil diameter was recorded throughout the session to assess the perceptual, cognitive, and response-related processing demands of the different labels and label sets. Similar to findings on error rate and time, the size of the pupil was not affected by label design or level of label similarity, nor was any interaction found between the two factors ($p > 0.05$). This indicates that the new template design or the between-drug similarity that the template introduced did not increase the workload of the pharmacists, keeping it consistent (3.53 mm on average) across the different drawers.

The fixation data, however, revealed some interesting findings. Although there was no main effect of similarity and no interaction between design and similarity for either fixation count or fixation duration ($p > 0.05$), a main effect of label design was found for fixation count ($p < 0.05$). The new template required *fewer* fixations (0.88 fixation per Abbott label per trial) than one of the existing designs, E3 (1.1 fixations). This suggests that the new design improved participants' search efficiency, which could be attributed to the key label elements (i.e., drug name and dosage strength) being closer together than in the existing designs. The closer proximity of the label elements created a smaller area that needed to be scanned to obtain the key information. In the new design, both elements were often processed with a single fixation, whereas the existing design sometimes required multiple fixations (Figure 10.20).

The fixation duration analysis also revealed a main effect for label design ($p < 0.0001$). However, this effect was counter to the number of fixations result; the new template required *longer* fixations ($M = 279$ ms) than existing design E3 ($M = 220$ ms). Increased fixation duration can be caused by unclear or ambiguous information, but because the drawers were equivalent in terms of content, this explanation was ruled out. A more plausible cause was the smaller font and higher

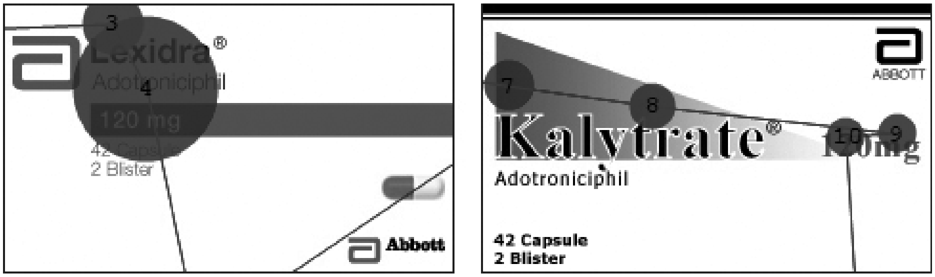


FIGURE 10.20

New label design (*left*) and one of the existing label designs, E3 (*right*), with superimposed sample fixation data. The circles indicate fixations, with the circle size being proportional to duration of fixation.

information density in the new template, which demanded more processing time per fixation. However, this result cannot be considered negative because it was accompanied by a decrease in the number of fixations, and time-on-task was not ultimately affected. It should also be noted that the font sizes in the new template were chosen to ensure both consistency across labels and readability of critical information at an estimated worst case distance of 24 inches (arm's length).

The purpose of this study was to determine the impact of the new template on pharmacists. Overall, we concluded that the new labels did no worse than the existing labels we tested. However, our previous study found significant performance improvements associated with the new template when it was applied to bottles. So why did the template not show much improvement when it was applied to cartons? One explanation that comes to mind involves the existing label designs to which we compared the new template. Some of the existing labels used in the previous study had very poor contrast and extremely high information density, and they were thus not nearly as readable as the labels used in the present study. The results of both studies indicate that although the new design is not superior to every Abbott label design that exists, it is better than some and, in the worst case, equal to others.

We realize that just like any study, this one has its limitations. Although our participants were domain experts, they were novices with regard to the particular (fictitious) drugs we showed them. Therefore, our findings only generalize to novice participants in early stages of learning. It is however possible that the across-drug label similarity that will accompany the introduction of the template will affect expert performance, preventing experts from relying on their top-down processes when searching for drugs. Instead of being able to conduct an efficient search by discriminating labels parafoveally, they might have to examine them one by one if they are not distinctive enough. Label similarity may also affect learnability because it is difficult to memorize what each drug looks like if they all look similar.

Although the smaller font used in the new template did not have a detrimental impact on the pharmacists in this study, it might cause issues when the distance between the eye and the drawer increases and possibly exceeds 24 inches. This impact could be exacerbated in high-similarity conditions where it is difficult to see the name and dosage strength of the drug and there are no other cues to identify it from the distance.

One recommendation is to conduct research with longer sessions or multiple sessions over a period of time to assess learning and expert performance. Another is to extend the study to physical (rather than digital) drawers and cartons experienced under a variety of conditions (e.g., various viewing distances or different cognitive workload levels), which would improve the ecological validity of this research.

In this study, a great deal of attention was given to the selection, definition, and collection of measures. When selecting measures, we had to account for the fact that error rate might be insufficient. Understanding the relationship between easily observable performance measures and other less overt physiological metrics allowed us to investigate underlying cognitive processes to assess performance.

Next, we defined the measures and determined how they would be analyzed. For example, deciding to collect the number of fixations was not enough. We had to define a fixation and choose whether to analyze this measure for the entire drawer, for all Abbott labels, or just for some of them. Recognizing factors that may impact these measures (e.g., targets attract more fixations, pupillary response has a delay, or mouse clicks create additional fixations) helped us refine these measures and adjust procedures and analyses to ensure that we were not collecting just data but valid data.

10.5.7 Biography

Agnieszka (Aga) Bojko is a chief scientist and associate director at User Centric, Inc. She directs global user research studies and publishes methodology papers that focus on integrating behavioral measures with eye movement metrics. Aga holds graduate degrees in Human Factors and Human–Computer Interaction from University of Illinois and DePaul University, respectively.

10.5.8 References

Bojko, A., Gaddy, C. Lew, G. S., Quinn, A., and E. Israelski. (2005). Evaluation of drug label designs using eye tracking. *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society,* Orlando, FL.

Fitts, P. M., Jones, R. E., and Milton, J. L. (1950). Eye movements of aircraft pilots during instrument landing approaches. *Aeronautical Engineering Review*, *9*(2), 24–29.

Flynn, E. A., Barker, K. N, and Carnahan, B. J. (2003). National observational study of prescription dispensing accuracy and safety in 50 pharmacies. *Journal of the American Pharmaceutical Association*, *43*(2), 191–200.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kotval, X. P., and Goldberg, J. H. (1998). Eye movements and interface components grouping: An evaluation method. *Proceedings of the 42nd Annual Meeting of the Human Factors and Ergonomics Society,* Chicago.

Salvucci, D. D., and Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocol. *Proceedings of the Eye Tracking Research and Applications Symposium,* Palm Beach Gardens, FL.

10.6 MAKING METRICS MATTER—By Todd Zazelenchuk

Collecting usability and other design-related metrics has become a hot topic in recent years, as usability has become more of a mainstream concept for many organizations. The consumer software industry, the world of home appliance design, and institutions of higher education are just a few examples of where organizational leaders have found themselves enamored with the collection of metrics as a way of helping their organizations "move the needle" in their product design efforts. Collecting quantitative measures of a product's performance, however, is only part of the equation. In order for usability metrics to stand a chance of influencing the future direction of a product, several criteria must be met. Without them, the effort may resemble a successful academic exercise, but it will most likely fail to have the desired impact on the product's direction. The following case study illustrates one such example where usability metrics were successfully collected, but their ultimate impact was limited.

10.6.1 OneStart: Indiana University's Enterprise Portal Project

Indiana University (IU) embarked on its enterprise portal project in the year 2000 with design research and iterative prototype development leading the way. Technically, the project had begun two years earlier with the publication of an information technology strategic plan for the university (McRobbie, 1998). This plan identified a broadening base of information consumers who were becoming increasingly tech-savvy and whose expectations for convenient, quick access to information and services were rapidly expanding. Although the plan never actually mentioned the word *portal*, it effectively described the need for what would become *OneStart*, a "next-generation" enterprise portal responsible for providing a full range of university services to over 500,000 students, staff, faculty, and alumni (Thomas, 2003).

Integral to the IU Strategic Plan was *Action 44*, the requirement for a user-centered design approach to all information technology projects. From 1995 to 2003, Usability Consulting Services, an internal consulting group based within IU's University Information Technology Services (UITS), supported project teams in the design and evaluation of their numerous software development initiatives. Known as the User Experience Group today, this team has since contributed significantly to the successful technologies delivered by UITS and Indiana University. In the case of the OneStart project alone, more than a dozen research studies have been conducted on various aspects of the portal over the past seven years, including usability testing, user surveys, and focus groups.

In 2000, not yet able to test any designs of its own, the OneStart team began with a comparative evaluation of some existing web-based portals. Three portals (MyExcite, MyFidelity, and MyYahoo) were evaluated with a sample of student and faculty users. The emphasis was largely on navigation and personalization tasks (selecting content for display, arranging a custom layout, changing background themes and colors). From this study, the team gained insights into many of the design elements that made portals of that era either easy or difficult for users to interact with and comprehend.

By early 2001, the team had a working prototype of OneStart in place, and the next phase of testing began. There were several motivations for the next round of research. At the most basic level, the team wished to understand how users would react to their university information and services being consolidated into the new portal environment. We anticipated that users may be confused about the relationship between the new portal and the traditional homepage of the IU website. Further motivation involved a desire to learn whether the content organization and personalization features of the portal were both usable and useful for the target population of users. Finally, the author was selfishly motivated to complete his dissertation related to the topic of measuring satisfaction as an attribute of usability. The combination of these motivating factors led to an empirical study with the following goals:

- Identify the major usability problems associated with the portal's navigation and personalization features in order to help direct the next iteration of OneStart.
- Establish usability benchmark data (comprising effectiveness, efficiency, and satisfaction metrics) for the core tasks currently supported by the portal in order to allow comparison with future design iterations of OneStart.
- Investigate the theoretical questions of whether certain methods of administering user satisfaction surveys have an impact on the ratings themselves and whether correlations between efficiency, effectiveness, and satisfaction exist for portal users.
- Identify why users rate their satisfaction with the portal the way they do (i.e., what are the contributing factors of a portal experience to users' satisfaction or frustration with the product?).

10.6.2 Designing and Conducting the Study

To achieve the goals outlined for the research, a usability lab study was designed and conducted with a sample of 45 participants representing the student portion of the overall OneStart target population. This was a much larger sample than the lab normally recruited for formative evaluation studies, but the desire to collect certain metrics and apply inferential statistical methods made it necessary. Had it not been for the dissertation-related questions, a smaller sample and descriptive statistics would have sufficed.

The study applied a between-subjects, one-variable, multiple-conditions design (Gall, Borg, & Gall, 1996), in which the 45 participants were distributed across three groups of 15, each of which encountered the same portal design and core tasks to be performed but experienced different conditions for rating their satisfaction levels with the product.

The tasks for each participant included a combination of information retrieval and personalization tasks. Information retrieval tasks consisted of locating ''channels,'' or groups of content to be added to the participants' portal page. Personalization tasks required the participant to change the look and organization of their interface (e.g., screen color, layout, add content, etc.).

A traditional two-room, mirrored glass lab facility was utilized with the researcher moderating the study from the test room, while the participant worked through assigned tasks in the participant room. The ISO definition of usability (ISO 9241-11, 1998), comprising the three attributes—*effectiveness*, *efficiency*, and *satisfaction*—was used as the basis for the metrics collected. For *effectiveness*, a rubric was established to judge whether task performances were scored as a pass or fail. A stopwatch was used to measure the attribute of *efficiency*, the time spent per task in minutes and seconds.

The third attribute, *satisfaction*, was collected using two different instruments: the After-Scenario Questionnaire (ASQ) and the Post-Satisfaction Survey of Usability Questionnaire (PSSUQ; Lewis, 1995). The ASQ consisted of three questions asked after the completion of each task. The PSSUQ consisted of 19 questions asked after the completion of the entire study. Both questionnaires utilized a 7-point scale (1 = Strongly Agree, 7 = Strongly Disagree) that was reversed prior to data analysis.

10.6.3 Analyzing and Interpreting the Results

We analyzed our qualitative data looking for high-frequency patterns of behavior that might suggest inherent problems with the design. We found several, along with problems that were lower frequency yet potentially severe in their impact on the user experience. Once this analysis was completed, we prioritized the problems based on frequency and our subjective ratings of severity to help prioritize the order of presentation in our final report.

The most frequently demonstrated problems involved personalization activities, with key problem areas including tasks such as creating a custom page for personal content, changing the color of a page, and viewing the completed page. These were all considered to be rather serious problems at the time, given the importance that the team believed personalization features would have on user adoption of the portal.

For the quantitative data collected, we calculated descriptive statistics for effectiveness, efficiency, and satisfaction to share with the project team. We evaluated effectiveness by calculating the mean values of task completion for each task, as well as the mean and standard deviation for all tasks combined (M = 0.731, SD = 0.238). Efficiency (mean time per task) was presented for

Table 10.10 Correlations between Usability Metrics

ISO Attributes of Usability	Correlations Found
Satisfaction effectiveness	(–0.593, $p < 0.01$)
Satisfaction and efficiency	(–0.452, $p < 0.01$)
Effectiveness and efficiency	(–0.394, $p < 0.01$)

individual tasks as well as for the full set of tasks (*M* = 467.4s, *SD* = 145.0s). Satisfaction was evaluated by reversing the scale values and computing the mean post-test PSSUQ scores for each group and for all participants (*M* = 5.1, *SD* = 1.1).

The theoretical questions for the study were analyzed further using SPSS to discover moderate to high correlations existing between effectiveness, efficiency, and satisfaction (see Table 10.10). The different satisfaction collection methods revealed no significant difference between methods (Zazelenchuk, 2002).

10.6.4 Sharing the Findings and Recommendations

The findings from the study were compiled and reported to the OneStart design team in both a written report and a presentation supplemented with video highlights of the most frequently occurring, highest-severity issues. Although this author has rarely compiled test session video highlights for presentation, this study represented an exception due to the large sample size. The impact of viewing tightly edited sequences of multiple participants (often ten or more) demonstrating the same unanticipated behaviors certainly drove the message home to the design team for many of the findings.

The quantitative data representing effectiveness and efficiency were shared with the design team on a per-task basis (see Figures 10.21 and 10.22). Given that there was no significant difference discovered between the three conditions applied in the study, users' satisfaction measures were presented as an average post-task score for all 45 participants.

From a practical perspective, the most actionable data collected from the study were the qualitative findings revealed in the prioritized problem lists and supported by the video excerpts in the summary presentation. A total of seven qualitative themes were identified representing participants' rationales for their ratings of satisfaction with the portal (Zazelenchuk & Boling, 2003) and in 2005 were part of Educause's recommended reading list for the Top Ten Issues in Information Technology (Educause, 2005).

The quantitative metrics were also shared with the design team, but a reliable frame of reference for their interpretation was lacking. Had the initial competitive evaluation of existing portals been conducted with the goal of establishing benchmarks for certain tasks, those results could potentially have represented a

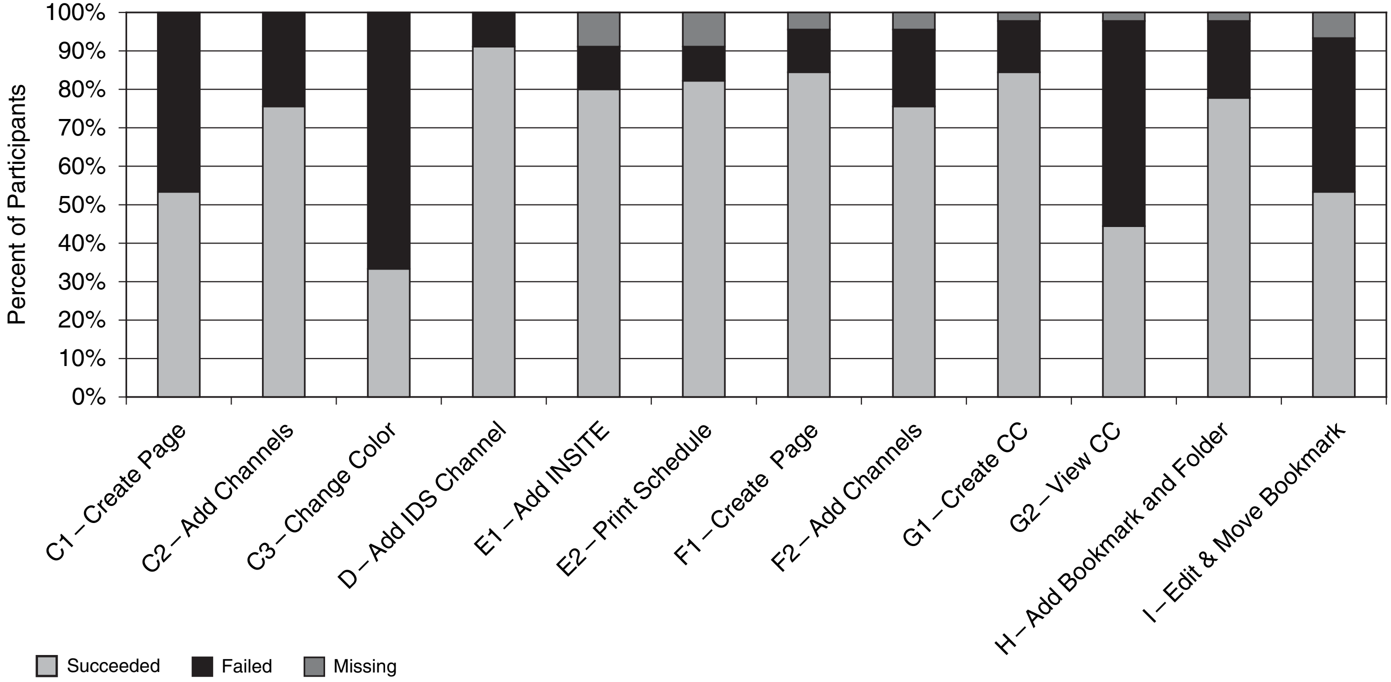


FIGURE 10.21

Task success and failure rates.

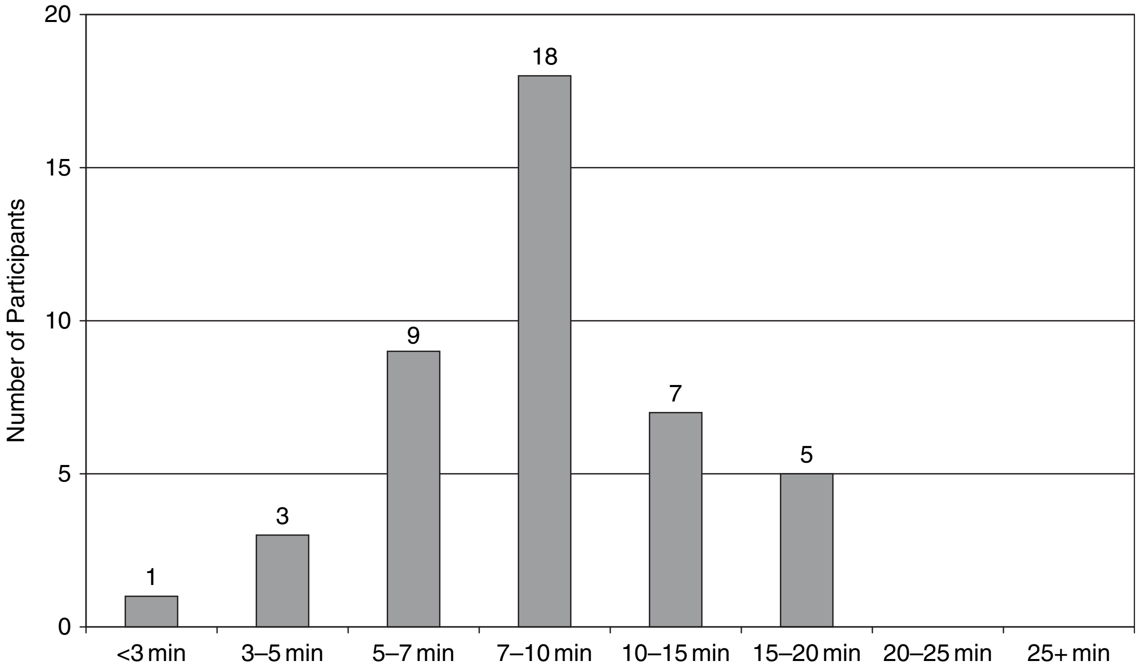


FIGURE 10.22

Mean time per task.

meaningful frame of reference for the analysis. Without those baseline scores, however, the metrics collected in this study were limited to answering the academic questions associated with the author's dissertation and providing an initial benchmark for future evaluations of the portal.

10.6.5 Reflecting on the Impact

Six years after the original study, and four years after the author's last direct experience with OneStart, an update from the design team provided additional insights into the challenges associated with making usability metrics matter. The metrics collected in the 2001 study had provided negligible long-term value. Although they successfully addressed the academic questions associated with the original study, their practical impact on the actual product was low. There were two primary reasons for this: Both represent important considerations for today's organizations as they race to institute a metrics-driven usability process.

Usability metrics only provide value when there is a frame of reference. Without it, teams are left to wonder whether 80 percent task completion is a good score, if 85 percent may be necessary, or just how much of a problem it is when someone "takes 30 seconds to locate the popcorn command the first time they use a microwave oven." When there is a clear plan in place for reliable, repeated measures to be collected in the future, an effective frame of reference can be established, and valuable comparisons and learning may begin.

In the case of OneStart, the metrics collected in the 2001 study represented the first attempt at measuring the usability of the portal. As a result, the numbers lacked a meaningful reference point and were much less actionable than the qualitative findings from the study.

Usability metrics are most reliable and informative when the tasks being measured represent core tasks that will likely persist throughout the life of the product. Spending time collecting metrics on anything but a product's core tasks contributes to the "frame of reference" problem by constantly measuring new tasks for the first time.

In the case of the original OneStart study, the tasks measured were largely feature-driven. In other words, they represented the tasks that the portal supported at that time, rather than those that were truly core tasks for the product over the long term. Moreover, those feature tasks have since been found to be less important than once imagined. Web server log data (another valuable usability metric), representing the actual usage of OneStart's personalization features over the past four years, have revealed that only 1 percent of users have ever visited the portal's personalization features. This has helped lead the team to rethink their emphasis on personalization options in the latest 2007 release by scaling back personalization to focus on simplicity, clarity of organization, and navigation. Given this change in direction, it suggests that collecting repeated measures of the original personalization tasks would not have been the best use of their time.

10.6.6 Conclusion

The OneStart case study represents a common example of where the efforts expended to carefully measure a product exceeded the returns. It reminds us that collecting usability metrics should be kept in perspective; they are a means to an end, where the ''end'' is the improvement of your product or process. By ensuring that you have in place a frame of reference to help you interpret your metrics, and that you restrict your focus to core tasks that can be revisited in future evaluations, you are more likely to produce metrics that matter.

10.6.7 Acknowledgment

Thanks to Dr. Philip Hodgson, Dr. Helen Wight, James Thomas, and Nate Johnson for their critical feedback on earlier drafts of this section.

10.6.8 Biography

Todd Zazelenchuk is a user experience researcher at Intuit in Mountain View, CA. He earned his Ph.D. in Instructional Technology from Indiana University in 2002. Prior to the consumer software industry, Todd worked in academia (Indiana University) and consumer goods (Whirlpool Corporation), gaining insights in to both the value and the challenges of applying usability metrics to the product design process.

10.6.9 References

Educause. (2005). Recommended readings on the top-ten IT issues—*www.educause.edu/ir/library/pdf/ERM0566.pdf.*

Gall, M. D., Borg, W. R, and Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

International Standards Organization (ISO). (1998). ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs). Part 11—Guidance on usability, 22.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction*, *7*(1), 57–78.

McRobbie, M. (1998). *Architecture for the 21st century: An information technology strategic plan for Indiana University*. Bloomington: Indiana University.

Thomas, J. (2003). Indiana University's enterprise portal as a service delivery framework, in *Designing portals. Opportunities and challenges.* A. Jafari and M. Sheehan, Eds. Hershey, PA: Information Science Publishing.

Zazelenchuk, T. W. (2002). Measuring satisfaction in usability tests: A comparison of questionnaire administration methods and an investigation into users' rationales for satisfaction. Dissertation Abstracts International.

Zazelenchuk, T. W., and Boling, E. (2003). Considering user satisfaction in designing web-based portals. *Educause Quarterly*, *26*(1), 35–40.