Performance-Based Usability Testing: Metrics That Have the Greatest Impact for Improving a System's Usability

Robert W. Bailey[1,3], Cari A. Wolfson[2,3], Janice Nall[3], and Sanjay Koyani[4]

[1] Computer Psychology, Inc - Sandy, Utah
[2] Focus on U! - Tallahassee, Florida
[3] Centers for Disease Control and Prevention - Atlanta, Georgia
[4] Food and Drug Administration - Silver Spring, Maryland
bob@webusability.com, cariwolfson@usabilityfocus.com,
sanjay.koyani@fda.hhs.gov, jnall@cdc.gov

Abstract. Usability testing methods and results have evolved over the last 35 years. With new advancements being introduced every year, it is important to understand the present state of the field and opportunities for further improvement. This paper will detail the research-based methods and metrics which are being used to ensure that usability recommendations are data-driven and performance-based. By focusing on the types of usability metrics being captured during usability tests, we will attempt to illustrate how usability researchers can quantifiably measure the performance of a system, use these measurements to make meaningful changes, and subsequently illustrate the improvements in user effectiveness, efficiency and satisfaction.

Keywords: Usability testing, Usability metrics, Effectiveness, Efficiency, Satisfaction, FirstClick, Usability methods.

1 Evolution of Usability Testing Methods

Before looking at the state of usability testing today, it is important to note how usability testing methods have evolved over the course of the last 35 years. At the 1972 Human Factors Society annual conference, Bailey [1] presented a paper that described the process Bell Laboratories had been using to conduct usability testing. At the time, the methods were considered first generation usability testing, wherein participants were tested one at a time with a usability researcher sitting next to them to manually record success and time on task. There were no real-time observers unless they sat quietly in the room behind the participant. Test sessions were videotaped using one camera pointed at the participant's face, hands and keyboard.

In the years following, usability researchers began to conduct testing in test facilities, complete with a one-way mirror for observers. Typical usability testing consisted of one-hour test sessions in which participants would perform a series of tasks, while thinking aloud. These tests focused on users' abilities to successfully complete tasks, with little emphasis on users' efficiency in completing the tasks.

During the test sessions, participants were generally allowed to take as long as they needed to complete a scenario while the usability test facilitator observed. The

M. Kurosu (Ed.): Human Centered Design, HCII 2009, LNCS 5619, pp. 3–12, 2009.

facilitator typically recorded comments made by participants, as well as notes about the user's behavior, e.g., frequent use of the 'Back' button.

Much of this usability testing focused simply on determining if participants were able to complete the tasks. The resulting usability reports made suggestions for improvements based on these aspects and focused on many of the qualitative issues discovered during testing.

By today's standards the tests were 'soft', and the test sessions were difficult to replicate, making it almost impossible to conduct valid and meaningful retests. Some tests were so qualitative in nature, that they actually resembled a 'live' heuristic evaluation of the system and focused less on quantitative metrics regarding users' success and efficiency in using a system.

While we do not discount the importance of qualitative observations made by skilled usability practitioners during usability testing, this paper will attempt to illustrate the ways in which these observations can more accurately be quantified and standardized, resulting in higher-quality testing and recommendations, substantiated by meaningful usability metrics. Consistent with Tullis and Albert [2], we attempt to capture the state of quantitative usability metrics that are now (or should be) included in current usability test reports.

2 Focus on Data-Driven Recommendations

Over the past five to ten years, the usability testing process has substantially changed and most likely will continue to change within the next few years.

One of the biggest shifts has been the emphasis on performance-based recommendations in lieu of more qualitative recommendations. In the past, many usability reports focused on recommendations based on the facilitators' observations and qualitative notes; today's usability reports use metrics to substantiate these observations and quantify the performance of a system.

These metrics are due, in part, to advances in technology that now automate much of the data recording and provide new levels of data that were not possible to capture manually. Sophisticated testing tools have been available for the past few years, and have substantially changed the way in which usability tests are conducted. Not only do these tools automatically capture much of the data, they also assist with the analysis of this data, considerably reducing much of the time previously spent calculating success rates, time on task, page views, etc.

The usability testing tool that we used to collect most of the data shown in this report is the Usability Testing Environment (UTE) [3] [4]. The Usability Testing Environment consists of two applications. The first is the UTE *Manager* which helps the usability researcher set up task scenarios (test-items), and pre-test and post-test questions. The UTE *Runner* then automatically administers the test to participants and tracks the actions of participants as they take the test, including clicks, keystrokes, and scrolling. Once the test is completed, the UTE Manager analyzes the results from all participants, and automatically produces a Word-based test report – complete with text, statistics and graphics – as well as an Excel spreadsheet with all of the raw data collected during the testing.

The Usability Testing Environment (UTE), and similar testing tools, have revolutionized usability testing for Web sites and Web applications. UTE has substantially reduced the time required to construct and conduct usability tests, and has improved the usefulness of test results.

3 Using Metrics to Substantiate Usability Recommendations

To develop performance-based usability recommendations, meaningful usability metrics must be consistently captured across all participants. We will share some of the metrics that we have found to have the greatest impact on improving Web sites and Web applications.

4 Important Sets of Data

There are three important, and very useful, sets of data generated by most modern usability tests. First is the performance data, which includes task scenario success, time to complete each scenario, and the number page views required to complete each scenario. Second are the preference data that are generated from questionnaires at the end of each scenario, and/or at the end of the test. Third, are the comments made by participants during the test, and their impressions gathered after they had completed all scenarios, including their overall impressions, and what they liked best and least.

4.1 Success Rates

Success rates are one of the most helpful sets of data associated with each scenario. There are two success rates that are fairly easy to collect, and are very useful. The first is the success rate with the initial or first click [5] [6]. The second is the overall success rate when totally completing each scenario.

The first click success helps us to determine whether or not the participants are starting on the 'right foot'. Whereas, the overall success rate provides an estimate of how successful users were in completing an entire task.

FirstClick© Analyses

One of the most significant and useful advances we have made in our usability testing is to focus on the user's first click in each scenario – particularly when participants are interacting with Web sites. Over the course of our testing, we have noticed that participants' ultimate success with a task was very closely related to what they did on the first page.

We analyzed users' first click success and ultimate success from various tests, across multiple Web sites and found that if the user's first click was correct, the chances of getting the overall scenario correct was .87. On the other hand, if the first click was incorrect, the chance of eventually getting the scenario correct was only .46, which is less than a 50-50 chance of being successful. In general, we found that participants were about twice as likely to succeed if they selected the correct response on the first screen. The correct/incorrect ratio was 1.9, with a range from 1.4 to 2.7.

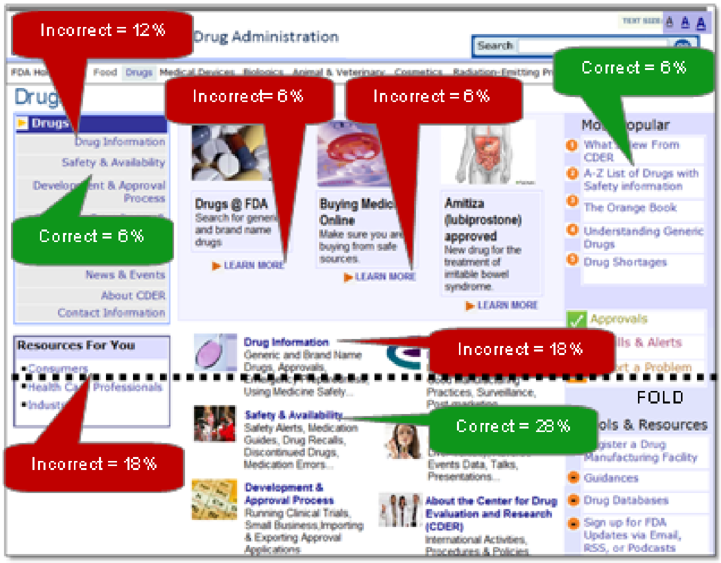

Fig. 1. FirstClick Callouts Shown on a Low-fidelity Wireframe for One Scenario

In presenting this data to design teams, we found that the best approach was to present the data scenario by scenario, and to include callouts to indicate where users clicked and whether the click would be considered a successful first click. To illustrate the data, correct clicks are shown in green, and the incorrect first clicks are in red, with a dotted line to show where the fold appeared during testing.

Ideally, the majority of participants would click on the correct link(s), and there would be few erroneous clicks. However, many times first clicks are made in a variety of different, and unexpected, locations. This data not only helps designers understand where users would look for information, but helps to validate an information architecture at the highest level of a Web site.

One major advantage of doing FirstClick testing is that far more scenarios can be included in a traditional one-hour test. This provides much greater 'task coverage'. Rather than using 10-15 scenarios during the traditional one-hour tests, we have been able to include over 100 FirstClick scenarios during the same period of time. This is significant, as previous studies [7] have found positive correlations between the number of tasks executed by participants and the proportion of usability issues found. In other words, the greater the number of tasks, the larger the number of usability issues identified.

Another advantage to FirstClick testing is the ability to uncover potential usability issues with a minimal number of prototypes and/or lower fidelity prototypes. Typical task scenarios require navigation through multiple pages, whereas, FirstClick tests only require the homepage and/or initial landing pages to be completed for testing.

Overall Task Completion/Success

The second success metric that can be captured is overall success, or simply whether or not a participant was able to successfully complete a task scenario within the given time limit.

In some instances, we are interested in measuring whether or not participants can find information on a site, whereas, in other cases, we want to see if participants can find the information and use that information to answer a multiple choice question correctly. Therefore, we judge success in one of two ways 1) when a user successfully navigates to the correct page or 2) when a user correctly answers a multiple choice question based on the content of a Web site.

For each scenario, success is either correct or not correct (binary). In our experience, success rates that are based on a facilitator's rating, such as correct, partially correct, and failure, can be very subjective and vary across usability practitioners. Therefore, we define success as either successful or not successful.

In interpreting success rates from usability reports, it is important to look at success in context of the scenarios asked. If the overall test-wide success rate is too high (80%-100%), the scenarios *may* have been too easy; if the overall, test-wide success rate is too low (< 50%) then the scenarios *may* have been too difficult.

A typical graph showing the success rate for each scenario is below.

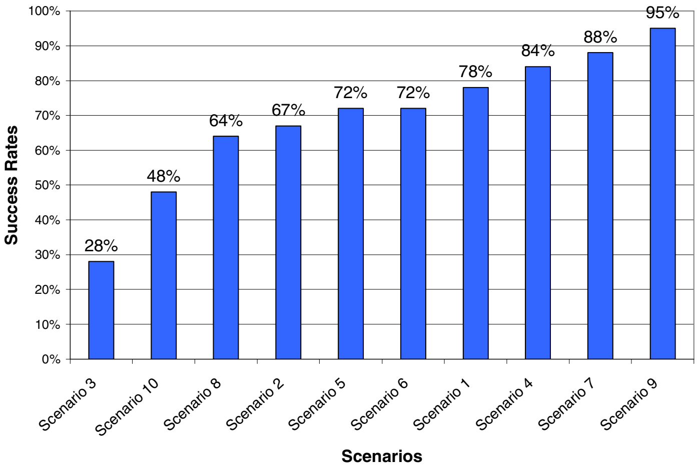

Fig. 2. Success Rates for Scenarios Presented in Order of Worst Performing

Once a test is completed, the scenarios are presented to the design team using a bar graph that shows the least successful scenario first, and the most successful scenario last. This helps to reinforce the idea that when making changes to the site, designers should start by fixing the scenarios that elicit the *worst performance*, and thus have the largest potential to make significant improvements.

4.2 Average Time to Complete Scenarios

The average time taken by participants to complete a scenario can be very informative and can help to measure users' ability to efficiently complete tasks in a reasonable amount of time.

When collecting this data, the average time to perform each scenario is usually measured from when participants see the first page (having already read the scenario), until they complete the scenario. Usually, shorter times indicate a well-designed site that allows users to efficiently complete tasks, whereas longer times *may* indicate that users had trouble with a task.

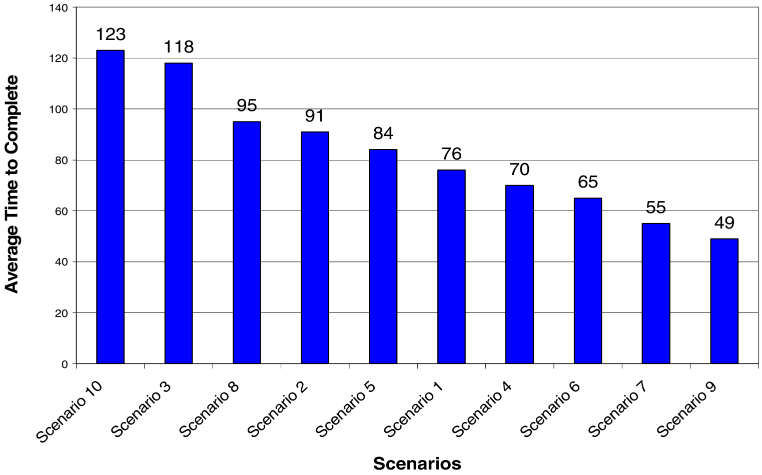

Fig. 3. Time on Task Presented in order of Worst Performing

When analyzing the results, we report the time on task for *successful* scenarios. This is to prevent the data from being eschewed by long times from participants who were not successful. It also allows design teams to determine whether or not the time to successfully complete a scenario is acceptable or needs improvement. A typical graph showing the average time in seconds is shown below.

4.3 Combining Success Rates and Average Time

Frequently, it is useful for designers to see the success rates and average time together. This can be done by providing a graph that combines both the success rate and the average time. Many times, the least successful scenarios take the longest time to perform, and the most successful scenarios take the least amount of time to perform. This provides two good reasons for presenting the results of a usability test in order of the scenarios with the worst performance. This type of graph is shown below.

4.4 Average Number of Extra Page Views

A fourth metric frequently used to evaluate users' efficiency is the number of average page views viewed by participants per scenario. Usually, the fewest number of page views leads to the fastest performance. Over several tests, we calculated a correlation of .82 ($p<.0001$) between the number of page views and the time taken to successfully complete the scenarios. In other words, average page views and average time to complete a scenario are usually related.

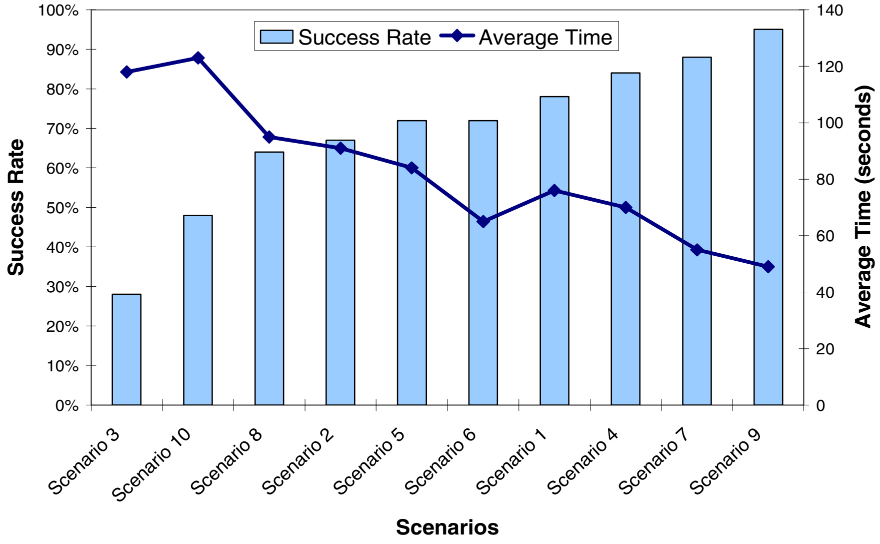

Fig. 4. Scenarios Presented in order of Lowest Success Rate overlaid with Average Time to Complete each Scenario

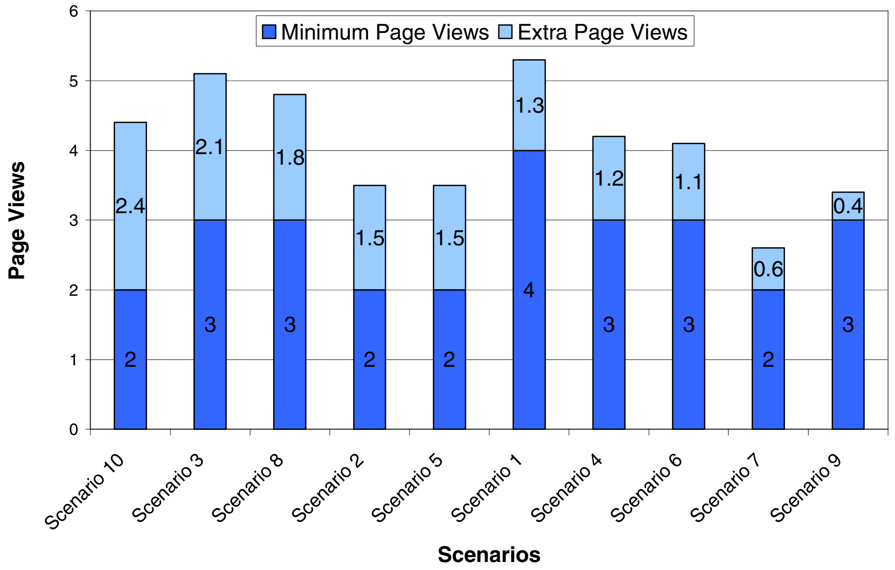

Fig. 5. Scenarios Presented in order of Worst Performing

While measuring the average number of page views per scenario is extremely valuable, we have found that a far more useful metric is a comparison between the average number of page views (for *successful* scenarios) with the minimum number of page views required to successfully complete the scenario (determined before the start of the usability test). By subtracting the minimum number of pages views from the average number of page views required to successfully complete the scenario, we obtain the number of extra or unnecessary page views.

When the number of extra page views is high, it suggests that some users were 'lost', had difficulty finding the information, or took a less than optimal path to find the content. When the number of extra page views is low, it suggests that users were able to efficiently find the information using an optimal path(s).

Following is an example of a chart used to present this data. The scenarios are presented in order of worst performing, with the scenario on the left having the greatest number of 'extra' page views (requiring more than twice as many page views than needed). The scenario on the right requires the least number of extra page views.

4.5 Combining Success, Average Time and Page Views

The previous discussion presents success, average time and average page views as separate metrics for analysis. However, Jeff Sauro [8] provides a way that usability specialists can combine these data into a single score for each scenario. Sauro's Usability Scorecard takes raw usability metrics (e.g., success rate, time, and page views) and calculates confidence intervals, z-scores, and quality levels. It then graphs the results automatically. The primary advantage of using the Usability Scorecard is the built-in confidence intervals. The confidence intervals can be used to help both testers and designers understand their test results better. This has particular value when comparing test results against a previous test's results, or against a usability objective.

4.6 Clicksteam Analysis

Additional performance-based metrics include a thorough analysis of users' behavior with a system. Previsouly, this type of analysis was too time-consuming or labor-intensive to undertake. However, with the emergence of new testing tools, we are able to easily log users' movements on a system, including clicks (right and left mouse clicks), text entry, scrolling and eye fixations.

For purposes of this paper, we have only included a brief summary of clickstream analysis. Clickstream analysis can be an extremely helpful tool in determining the effectiveness of a Web site's navigation. With a clickstream analysis, we can analyze the time spent on each page, the number of times users landed on a particular page, the number of times the 'back' button was clicked, etc. This type of analysis can help us determine whether users can effectively use a navigation system or whether users struggle on some tasks more than others. For instance, a task where users relied heavily on the 'back' button can tell us that the navigation schema was confusing or that a link label was misleading. We can also review the paths users took and review the number of times users viewed a page, such as a site map, search results, or even a 'help' file. This information helps us, as usability professionals, to better and more accurately quantify usability issues.

4.7 Preference Metrics

In addition to some of the above performance metrics, we also supplement our usability analysis with qualitative or preference metrics, including users' comments, concerns, frustrations and suggestions for improvement. In some instances, we evaluate or code user responses into similar categories in order to determine which elements users were commenting on the most.

Where possible, we also try to quantify users' thoughts and opinions with the use of post-scenario questions, as well as post-test satisfaction surveys. For instance, we have very successfully paired each scenario in a FirstClick test with a post-scenario question that asks users "How confident are you that the page you selected would have the information you are looking for?" This type of question allows us to analyze users' success in making a first click, the time it took users to select an item, and users' confidence that they would be able to find the information. By pairing confidence, satisfaction and preference data with performace data, we are able to have a more complete view of a system's usability, and thus, the recommendations for improvement that will have the greatest impact on the usability of that system.

5 Conclusions

By quantifying users' performance with a system, usability professionals can focus on the recommendations that provide the greatest potential for improvement, as well as ensure that we can reliably measure the effectiveness of these improvements in subsequent testing.

To that end, usability practitioners should try to focus first on scenarios with the:

- Lowest success rate
- Slowest average time
- Most extra page views
- Lowest efficiency, and
- Lowest satisfaction rate.

Caution must be exercised when usability reports are based strictly on the opinions of usability researchers or qualitative data gathered from participants' comments.

As practitioners, we must continually strive to quantify and justify the quality of our work and our usability recommendations. It is critical that we avoid making recommendations simply based on our opinions, or a 'live' heuristic review of the site, without clearly identifying those recommendations within our reports. One way that we have used is to relate each recommendation to a specific guideline in the *Research-Based Web Design & Usability Guidelines* book [9]. By working to quantify usability issues, we can ensure that recommendations are data-driven and performance-based, thus increasing the probability that improvements will measurably improve users' effectiveness, efficiency and satisfaction with a system.

References

1. Bailey, R.W.: Testing manual procedures in computer-based business information systems. In: Proceedings of the 16th Annual Meeting of the Human Factors Society, pp. 395–401 (1972)
2. Tullis, T.S., Albert, B.: Measuring the User Experience. Morgan Kaufmann, Boston (2008)

3. Bailey, R.W., Bailey, K.N.: Bailey's 'Usability Testing Environment'. In: Proceedings of the Human-Computer Interaction International Conference, July 22-27 (2005)
4. The Usability Testing Environment, `http://www.mindd.com`
5. Wolfson, C.A., Bailey, R.W., Nall, J., Koyani, S.: Contextual card sorting (or 'FirstClick' testing): A new methodology for validating information architectures. In: Proceedings of the UPA (2008)
6. Bailey, R.W., Wolfson, C.A., Nall, J.: Revising a Homepage: Applying Usability Methods that Guarantee Success. In: Proceedings of the UPA (2008)
7. Lindgaard, G., Chattratichart, J.: Usability testing: What have we overlooked? In: CHI 2007 Proceedings (2007)
8. Sauro, J.: The Usability Scoreboard, `http://www.measuringusability.com/scorecard/login.php`
9. Koyani, S.J., Bailey, R.W., Nall, J.R.: Research-Based Web Design & Usability Guidelines, U.S. Government Printing Office (2002, 2004, 2006)