Adaptive Usability Evaluation of Complex Web Sites: How Many Tasks?

Theodore Frick
School of Education

Michele Elder
School of Journalism

Christopher Hebb
School of Education

Ying Wang
School of Education

Sangil Yoon
School of Education

Indiana University
W.W. Wright Education 2276
201 N. Rose Ave.
Bloomington, IN 47405-1006

Abstract

Usability testing of a Web site is commonly conducted as a formative evaluation methodology in order to identify significant usability problems that need to be fixed. Usability testing can also be used to summatively evaluate the effectiveness of a Web site, i.e., to determine if it is working well for the target audience. Some debate exists as to how many tasks are needed to make a decision regarding the effectiveness of a Web site, and the question is raised: Can the number of tasks be adapted based on user performance during a usability test? This is particularly significant for Web sites which are very large and complex, and which would require a large number of usability tasks to evaluate them thoroughly. A Bayesian decision model, the Sequential Probability Ratio Test (SPRT), was investigated as an adaptive method for determining how many tasks are needed during usability evaluation of a complex informational Web site.

Twenty-five undergraduate students in the Indiana University School of Education were tested to determine if a part of the Web site was working well for teacher education students. Each user was tested with 20 randomly selected tasks during a one hour session. The SPRT was applied retrospectively to decide when testing of each subject could have been stopped to reach a decision of whether the site is working well for that subject. Results show that the decisions reached using SPRT are highly consistent with the decisions reached by testing the users with all 20 tasks. When the SPRT was applied, users could have stopped a usability test with an average of 12 tasks. Moreover, a decision could be reached about Web site effectiveness with 5 randomly sampled users in our study. In a summative evaluation of a complex Web sites effectiveness, the SPRT increases the efficiency of the usability testing by utilizing only as many tasks and users as necessary to reach a confident decision. However, reducing the numbers of users and tasks may be counterproductive if one is interested in identifying usability problems. In that case, more users and tasks are likely needed during formative evaluation.

Introduction

A usability test of a Web site is commonly conducted to either detect problems requiring improvement or to determine the effectiveness of the Web site, i.e., if it is working well for the target audience. Problem detection is more likely to occur as a formative evaluation process during the development of a Web site since developers intend to correct problems identified through usability testing. Some researchers recommend that testing five users is enough to detect most of the problems of a Web site (Nielsen & Landauer, 1993); however, this recommendation has been criticized by other researchers (Spool & Schroeder, 2001; Woolrych & Cockton, 2001), who claim that the lack of homogeneity of users and tasks often require many more users in order to identify major problems.

On the other hand, a Bayesian decision model, the Sequential Probability Ratio Test (SPRT) that was originally developed by Wald (1947), has been employed successfully in usability tests to determine the effectiveness of a Web site as a summative evaluation process (Frick et al., 2003). Wald’s SPRT methods are well-known in statistics and in manufacturing quality control. The SPRT provides rules to stop testing and choose one of two discrete alternatives using *a priori* decision error rates. The purpose of this study is to determine if SPRT can be used as a valid tool to improve the efficiency of Web site usability testing by testing as few tasks as necessary to make a conclusion on the effectiveness of a complex informational Web site comprised of about 6,000 Web pages.

Literature Review

Usability has many similar definitions but one that has broad recognition with the U.S. Government is the definition provided by the US Department of Health and Human Services (HHS, 2004): “Usability is the measure of the quality of a user's experience when interacting with a product or system — whether a Web site, a software application, mobile technology, or any user-operated device.” (n.p.)

There are many other definitions and descriptions that describe the multiple attributes of the user experience in the literature, but the three most common as cited by the International Standards Organization (ISO 9241-11,1998) are:

1. Effectiveness – accuracy and completeness of achieving the goals
2. Efficiency – speed and resources expended in achieving the goals
3. Satisfaction – Does the target audience like using the system?

Since the 1980’s there has been a profound shift in attention to user needs, as noted by the emergence of usability testing and usability test facilities. Usability testing has improved project quality, accelerated project delivery and provided dramatic cost savings, capturing the interest of both designers and managers (Nielsen, 1993; Shneiderman & Plaisant, 2004). During this time many usability-laboratory advocates split off from their academic roots into their own small businesses and consulting practices or joined larger firms that had usability labs. This trend has created a consulting community that provides testing for hire and currently remains an active service.

With that growth and maturity has come a shift in usability procedures. Dumas & Redish (1999) point out in the preface to their revised edition of the shift to more informal testing of usability: little or no videotaping, not logging every action, and smaller groups for each test session. One reason they claim is that "more acceptance of the fact that the value of usability testing is in diagnosing problems rather than validating products" (p. xi) reinforcing that large numbers of test participants are not necessary to feel confident that enough problems have been identified. Yet there is still a need for summative evaluation of products to make decisions on when to release the product to users and the general public.

Due to the costs associated with usability testing, the practitioner community has focused on the efficiency of usability testing. Perhaps one of the first widely recognized approaches to improving usability testing efficiency was Jakob Nielsen's "Discount Usability Engineering" (Nielsen 1993), which advocated what he called "the good" usability methods which do not necessarily give perfect results versus the "best" which may result in no usability testing being performed at all. This method is based on four techniques:

- User and task observation
- Scenarios
- Simplified thinking aloud
- Heuristic evaluation

As noted by Shneiderman and Plaisant (2004), one of the more controversial aspects to the recommendation is use of only three to six test participants in a round of usability testing.

How Many Participants are Needed?

Much has been made in the Human Factors / Usability community in the last decade about the proper number of users needed to perform a usability test on a Web site. Much of the research in the 1990's suggested that five participants will yield 80-85 percent of the findings in a usability test (Nielsen 1992, 1994, 2000; Virzi 1990, 1992; Lewis 1994). These recommendations are based on the use of a Poisson binomial probability distribution (Nielsen & Landauer, 1993).

Two of the more popular usability testing guidebooks, the *Handbook of Usability Testing* (Rubin 1994) and *A Practical Guide to Usability Testing* (Dumas & Redish 1999) discuss how many users are needed in usability testing. Rubin noted in regard to sample sizes "to achieve generalizable results for a given target population… one may need to test 10-12 participants per condition to be on the safe side, a factor that might require one to test 40 or more participants to ensure statistically significant results." (p. 29) Rubin also noted that it is often inappropriate or impossible to use classical experimental design procedures to conduct usability tests in the fast paced development environment. Rubin supports the Virzi and Nielsen testing recommendations, but takes it one step further – recommending four participants per treatment group.

Dumas and Redish (1999), one of the more popular usability guides whose first edition was published in 1993, discussed how many people to include in usability testing as part of the whole usability testing process. Starting from a task analysis perspective, Dumas and Redish first suggested identifying user profiles and then selecting subgroups that will need to be tested. Characteristics that are important for those subgroups are then defined and the most critical characteristics in participants are selected for testing. These are used to determine the range of participants, which then has an impact upon how many participants need to be selected for usability testing.

When recommending an actual number of participants to use, Dumas and Redish (1999) cite the Nielsen & Molich (1990) along with the Virzi study (1992) that claim 3 to 5 users are enough. Dumas and Redish seem to hedge a bit by noting most of the major problems are found with 3 to 5 users but most usability tests are with 6 to 12 participants. Their recommendation of "3 participants for each subgroup is probably an absolute minimum" (p. 128) seems to indicate they feel the 3 to 5 recommendation is best applied to the sub-group level, although this is as close as they come to a specific recommendation pointing out that you have to balance time, money and information gained when performing usability tests.

More recently there have been challenges to the assumptions about 4 to 5 users being enough, (Caulton 2001; Woolrych and Cockton 2001; Spool 2001; Molich et al 2004). These studies consider the effect of sub-groups upon the homogeneity assumption, the severity of the problem and the frequency at which it might occur, and the effect of different evaluators. The latter two studies also point out potential issues that should be considered for large, commercial web sites including visiting different portions of large sites and potentially conducting different tasks. In larger Web sites the validity of the tasks chosen to test compared to the full site usability could be a greater variable than having the correct number of users.

The SPRT was developed by Abraham Wald and utilized by the U.S. government in World War II for quality control of weapons. Wald demonstrated that by making observations sequentially and applying decision rules after each observation, then about half as many observations on the average are required to make a decision versus conventional testing with predetermined fixed sample sizes. SPRT does not take into account item difficulty variability or chances of guessing like more complex models such as item response theory (IRT). However, SPRT does not require the large amount of testing in advance to determine item response functions that IRT models require (Frick, 1992).

Application of SPRT to Usability

Limited literature exists that addresses the number of users needed to conclude if a Web site is working well (also noted in Frick et al. 2003). Perhaps because the emphasis has been upon problem identification during development (more formative evaluation), there seems to be little research on providing evidence that a Web site is effective in allowing users to reach their goals (summative evaluation).

Frick may have been the first to apply SPRT to usability testing. A 2001 study (Frick et al.) looked into applying Wald's SPRT to Web site usability testing. By focusing on Web site effectiveness (more of a summative evaluation) instead of problem identification, the authors applied SPRT to see if the average sample size to make a decision on when to stop testing would be reduced.

This study took 31 subjects and asked them to perform 20 random tasks selected from a pool of over 330 tasks associated with the Indiana University Bloomington (IUB) Web site. Subjects were recruited on a proportional basis according to the population served by the IUB Web site. The SPRT parameters were set *a priori* at: success level=.90, failure level=.50, α error=.05, β error=.05 and then applied at the task level for all 31 participants post hoc.

Application of *post hoc* SPRT would have resulted in 30 of the 31 decisions to be the same as though all 20 tasks had been put into the SPRT formula. Since the expected agreement was 90 percent (1-(α+β)), the SPRT made fewer classification errors (3.2 percent) than the expected error rate of 10 percent.

Using the same SPRT parameters and applying at the user level, the decision on Web site effectiveness could have been made after only four subjects. The study concluded that the IUB Web site was not effective with the SPRT parameters used when using the "short" SPRT test (4 users) or when it was applied to all 31 users. The study also calculated that applying SPRT would only take 12-20 percent of the testing time versus what was defined as a typical usability test (10 subjects with 15 tasks). This number should be viewed cautiously however, for if the Discount Usability Engineering procedures suggested by Nielsen are used (five users) the advantage would only be slightly less.

A second study by Frick et al. (2004) looked at specifically applying SPRT to the number of users on the IU Library Web site to determine the effectiveness of its electronic card catalog system (IUCAT). A total of 51 people were selected through a stratified convenience sample and were asked to attempt two specific but representative tasks to find holdings listed in the catalog using the search function for IUCAT. The two tasks chosen took about 10-15 minutes total, a design chosen so that more user tests could be obtained for the analysis. Participant test results were randomized to avoid possible bias from the results. The results were analyzed retroactively to determine how many participants would be needed to make a decision on effectiveness. Then various SPRT parameters were changed to see the effect upon the number of subjects needed to reach a conclusion. Utilizing baseline SPRT factors of success level=.90, failure level=.60, α error=.05 and β error=.05 a decision that the search engine was effective could have been made after 12 subjects. This was the same conclusion that would have been reached within the results from all 51 participants had been used.

Next, the study looked into reducing the zone of indifference (success minus the failure level). As the zone of indifference was reduced, more subjects were needed to make a decision. Likewise when the alpha and beta error rates were reduced, more subjects were needed to make a decision in the Frick et al. (2004) study.

Research Questions

This study intended to answer the following research question: When conducting a usability test with random task selection, how does the SPRT-based method of decision-making compare to the results of conventional usability testing? In other words, if tasks are randomly selected from a large pool for usability testing, does the SPRT method of deciding when to stop and make a decision agree with the decision that would have been reached with all tasks? A secondary question was: Can the number of users also be adapted by use of the SPRT?

Significance of Study

Wald's SPRT was classified as a defense secret by the U.S. government during World War II due to its practical effectiveness in making product quality control decisions. By applying SPRT in usability tests, researchers do not have to predetermine how many tasks to test before the usability testing starts. Instead, decisions of when to stop testing will be reached after analyzing data accumulated from the usability testing. Often SPRT leads to quick, but highly accurate, decisions regarding the effectiveness of a product with fewer tasks to test. Therefore it may increase the efficiency of Web site usability testing.

Usability testing can be very costly in both money and time. If the SPRT proves to be a reliable method in usability testing for determination of a Web site's effectiveness, it may improve the efficiency of usability testing and lower related costs substantially.

Methods of Usability Testing

Usability testing methodology was used to evaluate the School of Education Web site (Dumas, 1999). This approach involved having authentic users perform authentic tasks using the system, while a facilitator guided the usability session. An evaluator recorded the users' actions and comments while tasks were completed. During the testing, users were asked to perform a think-aloud protocol to help the evaluators understand their behaviors and gain insight into the design of the Web site (Ericsson, 1993).

Sessions were performed on an individual basis with each session lasting approximately one hour. Following the last session, qualitative and quantitative data were analyzed and summarized, and problem areas were identified. Participants in this study were provided with five extra credit points from their class instructors.

Participants

Participants for the study were recruited by the researchers from School of Education classes. One of the target audiences is current Teacher Education students who are at the undergraduate level. The following demographic criteria were used as guidelines:

- Gender – approximately 75% female, 25% male
- Student status – undergraduate level students
- School of Education, Teacher Education majors

A sample of 25 participants was obtained from 36 volunteers from four undergraduate classes

Procedures

Participants were asked to read and sign Indiana University Human Subjects consent form which included a brief description of the usability session, the user's risks, benefits, confidentiality and the researchers' contact information. Users were then asked to fill out a brief survey which included demographic questions, such as university class standing and prior experiences with the School of Education Web site.

Tasks were completed one at a time and recorded as: 1) success, if the answer to the question was found on the Web site, or 2) failure due to expiration of 3-minute time limit, or 3) failure because the user gave up, or 4) failure, if the answer was not found even though the user thought she or he had done so. Following the session, the users were asked a series of post-session questions which helped capture their overall experience with the Web site: 1) Now that you have used the School of Education Web site, what are your overall impressions of it? 2) If you could pick one aspect of the site that you did not like, what would it be? 3) What is one feature about the site that you found positive and liked? The users' qualitative post-session answers helped the research team to better understand specific concerns users may have with the system in areas that were not tested in the session.

Sessions were performed using Internet Explorer 6.0 and the Windows XP operating system on an Intel Pentium IV computer in the School of Education IST usability lab.

Tasks

Each user was asked to complete 20 tasks that were randomly drawn from a pool of 120 tasks generated from a prior needs analysis study. The tasks selected were ones that were considered relevant to teacher education students, as compared with those for alumni, faculty or K-12 professionals. Since tasks were randomly selected, there were very few tasks that were the same for most users. Tasks also were created in order to test users interacting with specific features of the system. Users were asked to complete as many of the 20 tasks as possible within a one-hour user session. Each user was given three minutes to complete a task. This time limitation allowed a user enough time to attempt all 20 tasks within an hour.

Methods of SPRT

The task data from each subject were run through the Web Tool for Sequential Bayesian Decision Making (Frick, 2003). The settings for the Bayesian Decision Making tool were as follows:

- Alternative A: Web site is working well
- Alternative B: Web site is NOT working well
- Minimum proportion of success if Alternative A is true: .85
- Maximum proportion of success if Alternative B is true: .60
- Error rate for choosing Alternative B when A is really true = Alpha error = .05
- Error rate for choosing Alternative A when B is really true = Beta error = .05

The task data were used in two ways. First, the total number of successes and the total number of failures per subject were entered to determine if the Web site worked well, did not work well or if more tasks were needed to make a decision based on all of the tasks. Next, the usability test was reenacted by entering the outcome of each task into the SPRT Web tool one item at a time for a given subject. The SPRT tool would return a determination after each task indicating if alternative A or B could be chosen, or if more tasks were needed to make a decision. If more tasks were needed, the next task result was entered. If a decision was made, then the number of tasks entered to that point was recorded along with the breakdown of the number of tasks identified as successes and failures.

Tests were run comparing the SPRT result from the overall (all tasks) test to the adaptive (task-by-task) test to evaluate the consistency between the two methods. In addition, the mean success rates across all participants for both the overall test and adaptive test were compared.

Results

Twenty-five subjects participated in the usability tests, 22 females and three males. The female/male split was 88 percent and 12 percent respectively, which is approximately 11 percent higher in females than the population of undergraduates in the Indiana University Bloomington School of Education in 2003 (76.7 percent female/23.3 percent male) (Indiana University School of Education, 2004).

The subjects included 11 freshmen, 11 sophomores, two juniors and one subject who did not report class rank. The subjects' academic areas of interest included Art Education (3), Early Childhood Education (3), Elementary Education (7), Special Education (4) and other (7). When asked how often the subjects visited the School of Education Web site, the following responses were marked: Very often (0), Often (2), Seldom (5), Rarely (10), Never (7) and one participant did not answer the question. This indicates that most of the participants were unfamiliar with the Web site used in the usability testing.

Table 1 shows on agreement of decisions reached from all 20 randomly selected tasks for each user and the SPRT adaptive reenactments. The value for *kappa* is 0.74 when corrected for chance agreement, with a significance level of $p < .01$ (simple percentage agreement is 22/25 or 88 percent). This means there is substantial agreement that the same result is achieved by using decisions that were made using all 20 tasks and the subsets of those that were made after reenacting the test using the adaptive SPRT. In other words, decisions reached with a subset of the randomly selected 20 tasks for each user were largely the same as decisions reached with all tasks. In 16 cases (subjects), the conclusion was that the Web site was not working well (with a maximum success rate of .60 or lower), using either the adaptive SPRT method or by using all 20 randomly selected items. In two cases, the conclusion was that the Web site was working well (with a minimum success rate of .85 or higher), using either method. In four cases, no decision could be reached at the *a priori* error rates, after exhausting all 20 items, using either method. The three cases where the decision outcome differed was when SPRT concluded the site was not working well, but the conclusion after 20 items was that no decision could be reached. See Table 1.

Table 1. *Crosstabulation: Did the Web site work overall based on the adaptive SPRT? versus Did the Web site work well with all tasks?* (*N*=25 users, *kappa* = 0.74, *p* < 0.01)

		Did the Web site work well with all tasks?			
		No	Yes	No Decision	Total
Did the Web site work well based on adaptive SPRT?	No	16	0	3	19
	Yes	0	2	0	2
	No Decision	0	0	4	4
	Total	16	2	7	25

The researchers observed improvements in solving tasks in some participants as their tests progressed. This led the researchers to consider that subjects were becoming more acquainted with the Web site as usability tests proceeded. Therefore, the success rate for the first half and second half of each usability test was determined. Then these figures were compared as seen in Table 2. There was an increase from the first half success rate (.56) to the second half of the test success rate (.65), and the difference was about half a standard deviation. It does appear that subjects did learn where things were located on the Web site and how it is organized, since they did perform better during the second half of the tasks. This may account for the three no-decision outcomes when SPRT concluded the site was not working for those subjects. The performance of those three subjects improved enough that it was no longer possible to confidently choose between the two alternatives (Web site works or not). For the majority, however, the learning that occurred did not affect the overall conclusion of whether the Web site was working or not at the levels specified by the Web director (.85 vs. .60)

Table 2. *Comparison of users' task success rate between the first set of 10 tasks and the second set of 10 tasks*

	N	Minimum	Maximum	Mean	Std. Deviation
First random set of 10	25	0.22	0.90	0.56	0.19
Second random set of 10	25	0.30	1.00	0.65	0.19

Table 3. *Number of tasks needed to stop testing with the SPRT and choose an alternative*

Participant Number	**Number of tasks to stop testing with SPRT**
#1	6
#2	6
#3	no decision
#4	13
#5	16
#6	13
#7	14
#8	20
#9	9
#10	9
#11	5
#12	17
#13	5
#14	6
#15	6
#16	6
#17	no decision
#18	10
#19	13
#20	6

#21	17
#22	10
#23	no decision
#24	5
#25	no decision

In Table 3 the numbers of tasks required to reach an SPRT decision for each case are listed. It can be seen that decisions could be reached with as few as five or six tasks for some users, and in four other cases, no decision could be confidently reached with all 20 tasks. More tasks would have been needed for those users in order to reach a conclusion. Twenty tasks were not enough in those cases. Table 3 shows how the SPRT makes it possible to adapt the number of tasks based on the performance of users and the decision criteria established in advance.

From Table 3, it can be seen that an average of about 12 tasks were needed in our study, but the range was from 5 to more than 20 (but we stopped with those subjects since the test session would have gone beyond the hour allocated). If users are selected at random from those in Table 3, and the SPRT is applied after a decision is reached for that user, a total of 5 users would be needed to reach the decision that the Web site was not working as well as the Web Director had hoped (alternative B is chosen for the Web site). Thus, if the SPRT had actually been used in this study, we could have reached this conclusion with 5 users and an average of 12 tasks per user – and we would have reached the same overall conclusion. This is a significant reduction in time and effort in conducting such usability tests, when compared to 25 users and 20 tasks each.

Discussion

The SPRT requires the decision maker to clearly establish the criteria that will be used in making a decision. 1) What is the minimal success rate that is acceptable to decide that the Web site is working well? 2) What is the maximum success rate that is acceptable to decide that the Web site is not working as well? 3) What probabilities are tolerable for making erroneous conclusions? For example, had the success rates been 0.60 vs. 0.40 in this study, with the same alpha and beta levels for decision error rates, then the conclusion would likely be that the Web site was working *well* – just the opposite of that found in this study. In fact, success rates on most Web sites are often less than 50 percent, according to Nielsen (2001), but he estimates that success rates have risen at a rate of 2.5 percent per year in the past few years due to increased usability evaluation that appears to be occurring (Nielsen, 2003). Thus, from this perspective, student performance in our usability tasks with this Web site were better than success rates on most Web sites. Nonetheless, the success rate was not as high as the Web Director had desired.

Furthermore, the 120 tasks chosen for this usability test were those which were chosen from a much larger list of frequently asked questions determined from prior needs analyses. The 120 tasks selected were deemed to be largely relevant for the subgroup we tested (i.e., undergraduate teacher education majors). The overall Web site is designed to address other target audiences that include: prospective students, graduate students, faculty, staff, associate instructors, alumni of the School of Education, and K-12 professionals. The large bulk of the 6,000 pages on this site were not applicable to the tasks chosen. Most of the answers to the 120 tasks could be found on the Web sites for the Office of Teacher Education, which consist of approximately 560 Web pages or roughly 10 percent of the overall Web site. Thus, no inferences should be made concerning the effectiveness of the Website for these other audiences, or for the 90 percent of other Web pages that are part of the overall School of Education Web site.

Finally, it should be noted that users in our study began their usability session at the IU Bloomington home page. These users did *not* begin at the School of Education home page, and we did *not* tell them that the answers were on School of Education Web sites, nor on the Office of Teacher Education Web sites. Thus, the participants in our study, who were largely unfamiliar with School of Education Web sites, were in essence looking for “needles in a huge haystack”. Indiana University currently has in excess of one million Web pages in numerous sites, according to the person who manages the IU search engine (Percival, 2005, personal communication).

Implications for Further Research

It is a common goal of usability testing methodology to identify problems that need to be fixed. While applying SPRT may help an evaluator determine the overall effectiveness of a Web site, stopping a user session early may result in failure to identify specific usability problems. Ultimately, the goal of usability testing is to evaluate the current state of a system, detect problems, interpret the reasoning behind the problem, and diagnose the problem with a potential solution. With the utilization of SPRT in a usability study, specific problem detection may not occur. One of the authors has since explored this issue in greater depth in his dissertation (Hebb, 2005). In further analyses of the data from this study, he found that with 5 users and an average of 12 tasks that:

> …this number of users and tasks identified 25% of known usability problems. Likewise, the Poisson-based approach identified 35% of the problems with these same 5 users but with 20 tasks per user. For over 10 years, the prevailing rule of thumb has been that 4 to 5 subjects will identify the usability problems when evaluating websites. Results from this study indicated that more subjects may be needed for identifying 80-85% of website usability problems. Results also indicated that the SPRT is more cost-effective than the Poisson model if summative evaluation and not problem identification is the goal of usability testing. (p. vi)

These findings suggest that efficiency in terms of minimizing numbers of users or tasks in usability testing is counterproductive if the goal is to identify usability problems. In the Hebb (2005) study, he identified 83 unique problems with all 25 users and 20 randomly selected tasks per user. If 10 users are chosen at random, 52 unique problems are identified (63 percent), and if 5 users are chosen only 29 unique problems were uncovered (35 percent).

It appears from Hebb's data that if problem identification is the goal of usability testing, then the five-user rule that Nielsen and Landauer have suggested may badly underestimate the number of subjects needed to identify 80 percent of usability problems. More research is clearly needed here.

It does appear from the Frick et al. (2003) study that the SPRT can be effectively used when a Website has few usability problems. In that study, the search engine for the electronic card catalog was evaluated. This search engine has evolved, been improved, and has been in use for over 15 years (even before the Web came about). In other words, if most usability problems have been fixed in a product, then the SPRT can be an efficient method of verifying product effectiveness in summative evaluation. For formative evaluation, however, fewer numbers of subjects and tasks appear to be detrimental to problem identification.

References

Caulton, D. A. (2001). Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology, 20*(1), 1-7.

Dumas, J. S. and Redish, J. C. (1999). *A practical guide to usability testing* (revised edition). Exter, England: Intellect.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, MA: MIT Press.

Frick, T. W. (1992). Computer Adaptive Mastery Tests as Expert Systems. *Journal of Educational Computing Research, 5*(1), 187-213.

Frick, T. W. (2003). Web tool for sequential Bayesian decision making. Retrieved December 8, 2004 from http://www.indiana.edu/~tedfrick/decide/start.html.

Frick, T.W., Dodge, T., Liu, X. & Su, B. (2004). How many subjects are needed in a usability test to determine effectiveness of a Web site? Paper presented at the meeting of the Association for Educational Communication and Technology, Anaheim, CA. Available online at: http://education.indiana.edu/~frick/aect2003/frick_dodge_liu_su.pdf.

Frick, T.W., Lee, J., Park, Y. J. & Pascoe, S. (2001). How many subjects and how many tasks are enough for usability testing? Unpublished manuscript, Indiana University.

ISO 9241-11 (1998). Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability.

Indiana University School of Education (2004) Promoting diversity. Reterieved December 12, 2004 from http://www.indiana.edu/~ediverse/stuenrgen.html.

Hebb, C. L. (2005). Website usability evaluation using sequential analysis. Bloomington, IN: Indiana University Graduate School, Ph.D. dissertation.

Lewis, J.R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors*, 36, 368-378.

Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. (2004). Comparative usability evaluation, *Behaviour & Information Technology*, 23 (1), 65-74.

Nielsen, J. (1993). *Usability Engineering*. Cambridge, MA: Academic Press.

Nielsen, J. (1994). Guerilla HCI: Using discount usability engineering to penetrate the intimidation barrier. In R.G. Bias & D.J. Mayhew (Eds.), *Cost-justifying usability*. (pp. 242-272). Boston: Academic Press.

Nielsen, J. (2000). Why You Only Need to Test With 5 Users'. Alertbox, March 19, 2000. Retrieved September 12, 2004 from http://www.useit.com/alertbox/20000319.html.

Nielsen, J. (2001). Success Rate: The Simplest Usability Metric. Alertbox, February 18, 2001. Retrieved October 1, 2005 from http://www.useit.com/alertbox/20010218.html.

Nielsen, J. (2003). PR on Websites: Increasing Usability. Alertbox, March 10, 2003. Retrieved October 1, 2005 from http://www.useit.com/alertbox/20030310.html.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability Problems. Proceedings of INTERACHI’93 (pp. 206-213). Amsterdam, The Netherlands: ACM Press.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. Proceedings of the ACM CHI’92 (pp. 373-380). Amsterdam, The Netherlands: ACM Press.

Percival, P. (2005, Sept. 29). Personal communication, Indiana University Information Technology Services.

Rubin, J. (1994). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons, Inc. New York, NY, USA.

Shneiderman, B. & Plaisant, C. (2004). Designing the User Interface (4th ed.). Boston: Addison-Wesley.

Spool, J. & Schroeder, W. (2001). Testing web sites: Five users is nowhere near enough. Extended abstracts of CHI 2001, 285-286.

Virzi, R. (1990). Streamlining the design process: Running fewer subjects. Proceedings of the Human Factors Society 34th annual meeting, 1 (pp. 291-294). Orlando, FL.

Virzi, R. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors 34*, 457-486.

Wald, A. (1947). *Sequential analysis*. New York: Wiley & Sons, Inc.

Woolrych, A., & Cockton, G. (2001). Why and When Five Test Users aren’t Enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Eds.), Proceedings of IHM-HCI 2001 Conference: Vol. 2 (pp. 105-108). Toulous, France: Cépadèus Éditions. Retrieved April 2, 2004 from http://www.netraker.com/nrinfo/research/FiveUsers.pdf

U.S. Department of Health & Human Services (2004). What is Usability? Retrieved December 13, 2004 from http://www.usability.gov/basics/