Running head: How Many Subjects?

How Many Subjects Are Needed in a Usability Test to Determine Effectiveness of a Website?

Theodore W. Frick

Tyler Dodge

Xiaojing Liu

Bude Su

Department of Instructional Systems Technology

School of Education

Indiana University Bloomington

May 26, 2004

Abstract

How can one determine *efficiently* if a Website is working well? Relatively small numbers of the target audience are needed to improve a product during formative evaluation and usability testing as part of product development and revision cycles. However, during summative evaluation, how many subjects are needed to determine product effectiveness?

When investigating the number of subjects needed for usability tests, a Poisson probability model has been found to be a reasonable fit to extant data. However, this model was chosen based on the number of subjects needed to identify important usability problems with a product, *not* for determining its effectiveness. To determine if a Website is working well, we investigated the predictive validity of a discrete Bayesian decision model: the Sequential Probability Ratio Test (SPRT) originally developed by Wald (1947).

Fifty-one people representing a campus community participated in a usability test of the university library online catalog search tool, and the results were analyzed *post hoc* with SPRT re-enactments to simulate sequential decision making after testing each subject. Across a range of parameters, the Bayesian SPRT reached the same conclusion as reflected by the entire sample with many fewer subjects, utilizing typically small α and β error rates. The study provides evidence of the usefulness of the SPRT decision model in situations where determination of effectiveness is the goal (whether a product works well or not). The SPRT maximizes efficiency by testing only as many users as are necessary to reach a confident conclusion.

Introduction

In the last decade many ordinary people with little or no computer expertise have been drawn to the Web. Whether buying a book online, choosing which college to attend, or finding recipes that use zucchini, the Web is undoubtedly convenient for those who have access to computers and the Internet. Once users find a Website that looks promising for the task in mind, they try to complete it. If they do so quickly and easily, then the Website has achieved its purpose – users have achieved their goals. If they run into obstacles, they may quickly turn to another Website that will help them buy that book or find that recipe. Usability of Websites is paramount. If Websites are not usable, people can simply leave with a few clicks of the back button in their Web browser.

During product development, formative evaluation nowadays should include usability testing, the primary purpose of which is to uncover serious design problems that need to be fixed. Researchers and practitioners such as Dumas and Redish (1999), Krug (2000), and Nielsen (2000) recommend rapid prototyping and iterative rounds of usability tests with small numbers of subjects in each round of tests in order to improve a product's design in an efficient manner.

When investigating the number of subjects needed for a round of usability tests, a Poisson probability model has been found to be a reasonable fit to extant data (Nielsen & Landauer, 1993; Virzi, 1990, 1992). However, this model was chosen based on the number of subjects needed to identify important usability problems with a product, *not* for determining its effectiveness.

Historically, the term 'formative evaluation' has been used to denote activities to determine product worth during the design and development process; whereas 'summative evaluation' has been used for such activities near the end of development or after a product is completed (Seels & Richey, 1994). The purpose of summative evaluation is to determine the

extent to which the product achieves its goals. For example, Kirkpatrick (1994) refers to four levels of evaluation: 1) the reactions of the target audience (do they like it?); 2) their performance or behavior (are they successful?); 3) transfer or integration into their lives or workplaces (do they use it?); and 4) impact on the larger organization or society (does it make a difference?).

The purpose of the current study is to offer an approach to usability testing during summative evaluation that incorporates a Bayesian decision procedure called the Sequential Probability Ratio Test (SPRT) to determine product effectiveness (Wald, 1947). Can humans use the product successfully to achieve their goals? Is the design the human-computer interface effective in this regard? This is a different type of issue than that of improving the interface during the development process, asking instead, does the product work well? How many subjects are needed in the summative evaluation to answer this question? Five? Fifty? Five hundred?

Rather than testing with a predetermined sample size, SPRT analyzes the knowledge accumulating during testing to determine when to stop testing, significantly reducing the number of subjects required. Wald's sequential probability ratio test (SPRT) went beyond the work of Thomas Bayes, who was concerned about how decisions can be reached as evidence accumulates. Wald's SPRT provides rules for when to stop collecting evidence and reach a conclusion. The SPRT also indicates the likelihood that we would reach a wrong conclusion. Originally used for manufacturing quality control decisions, the SPRT was considered so important that it was classified as a defense secret by the U.S. government during World War II.

Usability

Usability of a software product or computer interface is a collection of attributes, some of which are easier to define and measure than others. Preece (1993) acknowledges this when

describing the relationship between Human-Computer Interaction (HCI) and usability; "The goals of HCI are to develop and improve systems that include computers so that users can carry out their tasks: safely, effectively, efficiently and enjoyably. These aspects are collectively known as usability." (p. 14)

The International Standards Organization (ISO) similarly defines usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO #9241-11). Other definitions in the literature vary somewhat from the ISO definition, including characteristics such as learnability and memorability (Nielsen 1993), flexibility (Shackel 1986), and utility (Shackel 1991), but the primary attributes of effectiveness, efficiency and satisfaction remain the core of most definitions (see Table 1).

Effectiveness, as one dimension of usability, is often measured by reduction of mistakes or errors that users make. The presumption is that such errors indicate problems in the design of a software product. During the development process, it is important it identify problems with a design and to correct them. Thus, problem detection is important during the formative evaluation process.

Problem Detection

Most of the literature in usability testing and numbers of subjects concerns problem detection, and a central tenet is that, given enough users and evaluators, most if not all of a product's usability problems may be uncovered. Of course, when ungainly numbers would be needed, a balance must struck between investment in usability testing and returns on investment,

that is, identified problems. Problem detection studies traditionally use the probabilistic Poisson model to determine the number of subjects needed (Nielsen & Landauer, 1993; Virzi, 1990, 1992).

$$Uncovered\ Problems = N(1 - (1 - \lambda)^n)$$

N: total number of usability problems in the design

λ: proportion of usability problems discovered while testing a single user

n: number of subjects

Given an accurate probability estimate, this simple formula provides a fairly good prediction of the number of subjects needed to determine a certain proportion of usability problems, though not the number of subjects needed to determine the product's effectiveness. Offering the first evidence supporting use of the model, Virzi (1992) found that observing four or five users would reveal 80% of a product's usability problems, but this estimate and a host of related issues have been actively debated over the last decade. The accumulation of literature relating to problem detection has raised doubts regarding the certainty of the "five users" rule, as well as bringing to light several previously unrecognized issues relating to usability testing, including the probability of error detection, the assumption of homogeneity among users, the inconsistency between evaluators, and the definition of the usability task.

The first central issue relates to the probability of detecting a problem during testing. An average value of between .30 and .40 was suggested by a number of studies (Nielsen & Landauer, 1993; Virzi, 1990, 1992) and, based on the cumulative binomial probability formula, led to the statement that testing only four or five users will uncover 80% of the usability problems. Indeed,

the diminishing returns after testing five users, a rule-of-thumb popularized in Nielsen's (2000) online *Alertbox*, continues to gain acceptance. While the rule holds true for probabilities in that range, other studies suggest that the actual probability of finding usability problems may be considerably lower (Lewis, 1994), with the result that usability testing would require a significantly greater number of users. For the *p* value of .16 that Lewis found, fully twice as many users would be needed to find 80% of the problems. Further, though Virzi asserted that the more severe problems would generally be identified before those of lesser import, Lewis found no such correlation; indeed, findings by Spool and Schroeder (2001) likewise challenge Virzi's claim, indicating that testing with a small number of users could be problematic for products with potentially hazardous problems.

Not only challenging the accepted sample size, concern over the probability levels of error detection has brought other issues to the discussion of usability testing. To begin, Caulton (2001) concluded that the assumption of homogeneity among users—the equal likelihood of all users to encounter all problems—not only accounts for the discrepancy between Lewis and his predecessors but compromises usability findings based on the assumption. Virzi's (1992) binomial model, Caulton explains, assumes homogeneity among the subjects, who "must be equally likely to encounter *all* problems" (p. 2). By introducing two classes of usability problems (common and rare) into the model, Caulton duplicates Lewis' (1994) findings that rare problems are not likely to be detected with only five subjects. Moreover, Caulton shows that heterogeneous subgroups likewise create the need for increased numbers of users to detect the same number of usability problems. Further, Caulton's conclusion accounts for the assumption by Virzi (1992), uncorroborated by Lewis (1994), that the probability of detecting a problem is positively correlated to the severity of the problem: "it is possible that *p* and severity *were* correlated in

Lewis' data, but that subgroups masked the correlation" (p. 6). In this way, the discrepancy between Virzi and Lewis may be explained, but only by introducing the complex issue of user group composition into usability testing.

The problems associated with the homogeneity assumption were also put forth by Woolrych and Cockton (2001), who, like Caulton (2001), challenged the validity of Nielsen and Landauer's (1993) formula supporting their claim that five users are enough to detect the majority of usability problems. First, through a discussion of statistical theory, the authors showed that the probability of errors being found may be much lower than is fixed in the formula. To demonstrate their claim, they cite Spool and Schroeder's (2001) study in which goal-oriented testing drove the probability much lower than Nielsen and Landauer's 31%. Then, citing their own study of heuristic evaluation, the authors show that the probability of error detection depends not only on the severity of the problem but on differences between users, the same issue explicated by Caulton.

Just as different users encounter different usability problems, so do different evaluators identify the problems inconsistently, a pattern referred to as the evaluator effect (Hertzum & Jacobsen, 2001; Jacobsen, Hertzum, & John, 1998). In these studies and others (Molich et al., 1998), it was found that even when employing similar evaluation methodologies to test the usability of identical products, evaluators differ in their assessment of which observations constitute usability problems. The subjective and inconsistent identification of problems, even when using such relatively strict usability evaluation methods as cognitive walkthroughs and think aloud procedures among experienced professionals, lead to inter-evaluator agreement as low as 5% to 65%. On the one hand, this suggests that testing with multiple evaluators will uncover more and more varied problems than with a single evaluator, and indeed, Jacobsen, Hertzum, & John (1998) note that "the effect of adding more evaluators to a usability test resembles the effect of

adding more users" (p. 256). On the other hand, the disparity among evaluators problematizes the "apparent reality of usability improvement achieved through iterative application of usability evaluation methods" (Lewis, 2001, p. 346).

In an article cited above, Spool and Schroeder (2001) reveal a fourth issue central to the question of the number of users, namely the definition of the usability task. In contrast to Nielsen and Landauer's (1993) testing with clearly defined tasks, or what Hudson (2001) calls "task-directed" testing, Spool and Schroeder allowed users to define their own goals, or "goal-directed" testing. That is, the five-user rule relates to situations in which all users engage in the same tasks of the product under evaluation, but when testing entails authentic users engaged in authentic tasks, the probabilities of error detection may be no higher than .16; at such low levels, the number of users Spool and Schroeder found necessary may range from around six to over thirty.

But What about Effectiveness?

Formative evaluation is not the same as summative evaluation. While iterative rounds of usability tests help identify problems with a product design and contribute to its improvement during the development process, such results do not imply that the product is *effective* in helping users accomplish their goals with the product. What is the likelihood that what is observed with a relatively small sample of users will be true of *all* users for which the product is intended?

A whole body of research in inferential statistics is concerned with problems of generalization from a sample to a population from which the sample is drawn (cf., Kirk, 1995; Fisher, 1956; Schmitt, 1969). When we make a generalization from a sample to a population, we

are making an inductive inference. Steiner (1988) states the form of inductive inference as follows:

1. *A* is true of $b_1, b_2, \ldots, b_n$; and
2. $b_1, b_2, \ldots, b_n$ are some members of class *B*;
3. hence, *A* is true of all members of class *B*. (p. 91)

In the present example, *A* would be the observations that indicate that the software product is effective; the *b*'s are the subjects in a usability test who represent the larger population of users, *B*. The inductive inference is that if the product is effective with *N* users in our sample, then it is likely to be effective in the overall population. The inference is also statistical, meaning that we are able to estimate the uncertainty or probability of making an error in such an inference. It is well known that the standard error in the sampling distribution of the mean is inversely proportional to the square root of *N*. Thus, if the standard deviation in the population is 30 (e.g., on a measure of effectiveness), and we tested 9 subjects, the standard error would be 30/3 = 10; where as if we tested 100 subjects it would be 30/10 = 3 (cf., Kirk, 1995, p. 51).

Suppose that we use success rate as our measure of effectiveness, and that we have a method of determining whether each subject is successfully able to perform authentic tasks in each usability test. If the *entire* population of users of a product is 100 and we test all 100, and 80 of them were successful in the usability test, then we know that the success rate is 80 percent. There is no uncertainty here, since we measured the entire population. However, if the population is 200 *million* users, and 80 are successful in our sample of 100, what is the likelihood that the success rate will be 80 percent in the entire population? We can construct a confidence interval based on the standard error, and we might make a statement such as we are 95 percent confident that the success rate is somewhere between 73 and 87 percent. On the other hand, if our sample size is 5

users, then we might be 95 percent confident that the success rate in the population is somewhere between 30 and 100 percent. The size of the confidence interval will depend on the standard deviation, but the conclusion is that the confidence interval will shrink as the sample size increases.

Alternatively, we could obtain a success rate for each user, for example, expressed as the percent of usability tasks completed successfully. Then we could compute a mean success rate and standard deviation for our sample. However, we are still faced with the same issue of sample size and confidence interval estimation.

Testing 5 to 10 users during a round of usability tests is likely to uncover major problems with a product assuming that appropriate subjects and tasks are used. However, we know from inferential statistics that, even if most of the subjects are highly successful in the usability test, we would have a fairly high degree of uncertainty in making an inference from such a relatively small sample to the population regarding the effectiveness of such a product. How many subjects do we need, then, in order to be more certain in our inductive inference?

While a considerable number of research studies address the identification of usability problems, our research team was unable to identify any significant literature addressing the number of users needed to conclude if a product is working well enough to stop further testing. Perhaps the simplest and most intuitive method is a simple calculation of success rate, or the percentage of successes encountered during usability testing. As Nielsen (2001) explains, success rates "provide a general picture of how [a product] supports users" and represent "the bottom line of usability" (n.p.), but beyond an explanation of the usefulness of tallying partial successes, he does not discuss such implications as the statistical limitations of such a metric.

Determining Effectiveness Using SPRT

The purpose of the current study was to investigate the Sequential Probability Ratio Test (SPRT) (Wald, 1947) as a method to determine the number of subjects needed to conclude whether or not a Website is effective. Wald's (1945) SPRT offers an elegant framework for making statistical decisions between two discrete alternatives when making sequential observations. Though not developed under the framework of Bayesian reasoning, SPRT can be regarded as an extension of Bayes' theorem with addition of stopping rules (Frick, 1989). Going beyond Bayes' concern about how decisions can be reached as evidence accumulates, SPRT analyzes the knowledge accumulating after observing or testing each subject in order to determine whether more usability tests are needed. The SPRT can reduce the number of subjects required when there is a clear pattern of evidence during early usability tests. The SPRT also tells us the likelihood that we would be reaching a wrong conclusion.

Wald (1947) claimed that using SPRT leads to an average saving of at least 48% in the necessary number of observations, compared with classical hypothesis testing with the same decision error rates. Later Colton and McPherson (1976) similarly found that using SPRT can achieve potential economy by testing fewer samples than with a fixed-sample-size while still maintaining the same α and β error rates.

The central tenet of Bayes' theorem is the likelihood principle: posterior probabilities are proportional to prior probabilities multiplied by the likelihoods of those alternatives. "Prior probabilities represent our knowledge *before* the current observation." (Schmitt, 1969, p. 260) A likelihood is the conditional probability of an event or observation given that a particular alternative is true, and a posterior probability is the conditional probability *after* the observation(s) (Schmitt, 1969, p. 83). For example, in the present context a likelihood would be stated as the

conditional probability of observing a successful subject if the site is working well – i.e., effective – in contrast to the likelihood of observing a successful subject in a usability test if the site is not working well – i.e., *p*(success | site is effective) vs. *p*(success | site is ineffective). A prior probability expresses our knowledge about the probability of an alternative *before* collecting new evidence. A prior probability can be entirely subjective, based on one's belief; or it can be based on empirical evidence previously gathered. A posterior probability of an alternative is the result of combining the prior probability with the new evidence (Schmitt, 1969). When observations are made sequentially, the posterior probability of one observation becomes the prior probability of the next observation. When two or more alternatives are involved, Bayes' theorem can be expressed as follows (Schmitt, 1969, p. 65, *e.g. added*):

If

i. Alternatives are mutually exclusive and exhaustive;

ii. Let $P_0(A_i)$ be the prior probability of A_i;

iii. X is the observation (e.g., subject succeeds, or subject fails the usability test);

iv. $P(X \mid A_i)$ is the probability of the observation given that A_i is true.

Then the posterior probability of A_i is

$$P(A_i \mid X) = \frac{P_0(A_i)\, P(X \mid A_i)}{\sum_j P_0(A_j)\, P(X \mid A_j)} \qquad (1)$$

Assuming we want to decide between two alternatives with error probabilities of α = *p*(choose alternative II | alternative I is actually true) and β = *p*(choose alternative I | alternative II is actually true) , Wald proved mathematically that the three rules below will yield decisions that will be wrong no more often than specified in the α and β error rates:

Rule 1: Compute the ratio (*PR*) of the posterior probabilities of the alternatives. If *PR* is greater than or equal to $(1 - \beta)/\alpha$, then choose the first alternative;

Rule 2: If *PR* is less than or equal to $\beta/(1 - \alpha)$, then choose the second alternative;

Rule 3: If neither Rule 1 nor Rule 2 is true, then another observation is needed. After a new result is obtained, then update the posterior probabilities and reapply the three rules.

Let us apply this reasoning in the context of the present study: we want to decide between the alternatives that either the Website is effective (I) or the Website is not effective (II). In this case, we would conduct one usability test at a time, and then apply Wald's stopping rules after each test in order to determine which option to choose. After randomly selecting a subject using the web site, we calculate the probability ratio, *PR*:

$$PR = \frac{P_{e0}\, P_e{}^{s} (1 - P_e)^{f}}{P_{n0}\, P_n{}^{s} (1 - P_n)^{f}} \qquad (2)$$

In Formula 2, *s* refers to the number of subjects who have successfully completed their tasks on the Website, and *f* refers to the number of subjects who failed to complete their tasks. P_{e0} and P_{n0} are the initial *prior* probabilities of effectiveness and non-effectiveness of the Website. These probabilities can be subjective priors if no empirical evidence is available, or they can be estimated from extant empirical data. Or if unknown, they can be set to 0.50 each, which then drops them from the equation (and hence becomes Wald's original formulation).

P_e is determined by the decision maker or investigator. What is the desired minimum success rate, if the site is working well? This is the lowest probability of success that is acceptable. Will the decision maker be pleased if 80 percent or more of the users can complete their tasks satisfactorily? Or must at least 99 percent of the users succeed?

Similarly, P_n represents the probability of success when the site is not working well. P_n is also determined by the decision maker. This is the maximum probability of success that is still considered to be unacceptable to the decision maker. For example, would the decision maker be unhappy if 60 percent or less of users in the population were successful?

P_n must be less than P_e; and the gap between them Wald referred to as the "zone of indifference." This gap may be puzzling initially, but the fact is that unless we sample the *entire population of users*, we will need to tolerate some uncertainty in the estimates of success rate. For a more familiar example, consider the results from surveys such as Gallup which often report an error rate for their obtained measure (e.g., plus or minus 3 percent). This usually means that if we had observed, for example, that 43 percent of those surveyed approved of the President's performance as Chief Executive, then we could be 95 percent confident that the actual percentage in the population (which we could not practically observe) was somewhere between 40 and 46 percent, at the time of the survey. We cannot be any more precise than that, based on the sample size (often around 1,000 subjects). In this example, we are pretty sure that the approval rating is not less than 40 percent, or not more than 46 percent. This is similar to Wald's "zone of indifference" – except however, we are trying to be confident in choosing between alternatives *outside* this zone to make a decision. For example, if the approval rating were less than 40 percent, then one might decide to change the political strategy; but if it were higher than 46 percent, then keep the strategy the same. In our context, if we were to conclude that the Website is not effective, then we would do something to fix the design; otherwise if it is working well, we would maintain the current design. Thus, Wald is requiring the decision maker to specify in advance the range and location of the zone of indifference, and the SPRT will tell the decision maker when the data indicate that the population is likely to be outside that zone, either above it or below it. If this kind

of thinking is difficult to do, and the decision maker would rather specify a single cut-off point, such as 0.85, then he or she also needs to specify a confidence band (e.g., plus or minus 5 percent). Then, in this case the upper and lower bounds would be 0.90 and 0.80, respectively, for P_e and P_n.

In order to make valid inferences, Wald's SPRT requires that:

1. Observations are independent. This means that the outcome of one observation should not influence the outcome of another. This assumption is normally met by conducting individual usability tests, one subject at a time and by not letting subjects help each other or communicate with each other about the Website or usability test prior to testing.
2. Observations are randomly sampled. Random sampling is necessary for generalizing results from a sample to a population. This assumption is harder to meet in practice, since we often do not have the time or money to do true random sampling as is often done in polls such as Gallup. We want to test subjects who represent the target audience in terms of their demographics. Perhaps the best strategy is to select first a relatively larger pool (e.g., 50 – 100 representative subjects, but nonetheless a convenience sample), and then randomly select them from that pool one at a time to conduct a usability test.

After each observation is made, *PR* is recalculated using the values of P_e, P_n, and the number of successes, *s*, and failures, *f*, using Formula 2 above. If *PR* is greater than or equal to $(1 - \beta) / \alpha$, then it is not necessary to make further observations; and the conclusion is that the Website is effective with a success rate of P_e or higher in the target population of users. Such a conclusion would be expected to be wrong no more often than the β rate. On the other hand, if *PR* is less than or equal to $\beta / (1 - \alpha)$, then we can conclude that the Website is not effective with a success rate of P_n or lower and an expected error rate no higher than α. If neither Rule 1 or 2 is true, then it is not possible to choose one alternative (that the success rate in the target population is P_e

or higher) vs. the other alternative (that the success rate in the target population is P_n or lower), given that the decision maker is unwilling to reach an incorrect conclusion at a rate higher than α or β. If Rule 3 is true, it means we cannot make a decision at the level of confidence that was initially specified, given the data collected (numbers of successes and failures) and the success rates of the two alternatives (effective vs. ineffective Website).

The reader should note that the observed success rate in the sample, $s/(s+f)$, is not being compared to P_e or P_n. For example, in Figure 1, it can be seen that $P_e = .85$ and $P_n = .60$, thus bracketing the zone of indifference. Notice in this example that three different success rates are illustrated: $X = 0.48$, $Y = 0.80$, and $Z = 0.87$. X is clearly in the shaded area for choosing alternative *II* (site not working well), but depending on the total number of observations, we cannot confidently choose alternative *II* unless *PR* is less than or equal to $\beta/(1 - \alpha)$. Similary, Z may be the observed success rate, yet we cannot choose alternative *I* (site working well) unless *PR* is greater than or equal to $(1 - \beta)/\alpha$. Finally, it may be the case that the observed success rate, Y, is in the zone of indifference, yet we could choose alternative *I* if *PR* is greater than or equal to $(1 - \beta)/\alpha$. This may seem puzzling to the reader, but one must remember that we are dealing with sampling error and attempting to make an inference about the population of users. In this case, the sampling error resulted in an observed success rate, Y = 0.80, which is outside the shaded area for alternative *I* in the target population. However, the observed sample is much more likely to be from the target population when *I* is true compared with *II*. How likely are we to be wrong? We would reach an incorrect conclusion at a rate no higher than β. That is what Wald proved mathematically (cf., Schmitt, 1969).

Another way to think about this is to estimate a confidence interval around an observed success rate in the sample of users (e.g., we might be 95 percent confident that the success rate is

somewhere between 0.73 and 0.87 in the target population). If the confidence interval includes one of the shaded regions in Figure 1 but not the other, then we would choose the respective alternative. Otherwise, the confidence interval would include parts of both shaded regions so we would not be able to choose between them. At some point, as our sample size increases and the confidence interval decreases (it is inversely proportional to the square root of *n-1*), we would expect to find that the confidence interval contains part of the region of one, but not both, of the alternatives. However, it might happen that with a very large sample, the confidence interval includes *neither* of the regions of the alternatives: that is, it is exclusively in the zone of indifference. Hence, we cannot choose between the alternatives. This is also one of the outcomes in inferential statistics after collecting data: drawing no conclusion.

There is a well-known problem in statistics when attempting to estimate a confidence interval for a proportion, as we are attempting to do here. The sampling distribution for proportions can be approximated by the normal (Gaussian) distribution when the success rates are not close to one or zero and the sample size is relatively large. Confidence intervals are based on *z* values from the normal distribution. For example, if $z = 1.96$, this contains 95 percent of the area in the normal distribution, the confidence interval would be plus or minus 1.96 standard deviation units above and below the observed success rate, as alluded to in the previous section. However, when the success rates are relatively high or low (close to one or zero), the sampling distribution is skewed, and when the observed sample size is relatively small, it is not appropriate to use the normal distribution as we usually do for estimating the confidence interval or standard error. Otherwise, we find ourselves making nonsensical statements such as we are 95 percent confident that the success rate in the population is somewhere between 87 and 112 percent (how could the

rate be higher than 100 percent?). See Hays (1973, pp. 378-380) or Ferguson (1971, pp. 143-144) for discussion of this problem.

The SPRT uses the binomial probability model in order to choose between two discrete alternatives. Wald showed that we could decrease the sample size by conducting the statistical test after each observation (hence the name *sequential* probability ratio test). When there is a clear trend early in the sampling, we can choose one alternative over the other with the same degree of confidence had a larger fixed sample been chosen and the statistical test applied after all the observations were complete. Wald showed that most of the time we could reduce the number of observations (by about half, in the long run). This was of immense practical significance when the tests were costly or destroyed the product being tested. For example, when testing bullets or bombs, the product would be destroyed and could never be used again. By reducing the sample size, if the sample were good, then more of the batch would be left intact for use.

Computational examples of how the SPRT works are given in Frick (1989; 1992). The reader can also experiment with a Web tool that Frick (2003) has since developed online at http://www.indiana.edu/~tedfrick/decide/start.html, in order to get a feeling for the working of the SPRT and the unfolding of the Bayesian decision process. Further examples are given in the results section that follows in this article.

The SPRT has been applied in industry to test the quality of manufactured products. In education, Bayesian procedures have been used in computerized adaptive testing (CAT) to make mastery and nonmastery decisions. For example, studies have shown that the SPRT can be used successfully during computer-adaptive mastery tests (Frick, 1989; 1992; Lewis & Sheenan, 1990; Reckase, 1994). Frick (1989) argued that though SPRT does not take into account variability in item difficulty, discrimination, and guessing factors, the decisions of mastery or nonmastery

reached by SPRT in his study agreed very highly with those reached through administering entire item pools to examinees. Frick concluded that because of its relative simplicity and practicality, the SPRT offers a viable model for mastery and nonmastery decisions, provided that the method is used conservatively (e.g., small alpha and beta error probabilities).

Similar to determining mastery or nonmastery of a educational objective, the task of determining site effectiveness is fundamentally a binary decision – either it is satisfactory, or we need to fix it; moreover, the task of determining site effectiveness with the fewest subjects possible is similar to the task of determining mastery by sampling as few test items as possible. It was this analogy that led us to the present study where we investigated whether the SPRT has predictive validity in reaching conclusions as to a Website's effectiveness using as few subjects as possible.

Method

Subjects

A total of 51 people 18 years or older participated in this study at a large mid-western university and its community. The subjects were recruited through a method of stratified convenience sampling. First, we identified five strata of users of university library resources: undergraduate students, graduate students, faculty, staff, and non-university affiliated community members. In order to have enough subjects for various retroactive SPRT analyses, we aimed for about 50 subjects, proportionally stratified by known demographics: 33 undergraduate students, 11 graduate students, 3 faculty, 2 staff, and 2 community members.

Among the research participants, 5 subjects reported that they used the university library's online catalog often, 18 subjects used it occasionally, 20 seldom used it , and 8 had never used it at all. In addition to self-reported usage of the online catalog, subjects were asked to report their

confidence using other similar search engines. Specifically, when asked to respond to the statement "I am confident using search engines" in terms of a 5-point Likert scale, 15 subjects strongly agreed, 27 subjects agreed, 7 subjects were neutral or undecided, 1 subject disagreed, and 1 subject strongly disagreed with the statement of confidence.

Task Selection

This research involved testing the usability of the search engine in the online catalog for the university library system (see Figure 2). While of interest to stakeholders in the site's usability, determining the success of this search engine remained of secondary interest. Our primary question focused on the SPRT for determining how many users were necessary to conclude whether or not the search engine was effective.

In order to provide some empirical basis for our task selection, we consulted documentation relating to usability testing of another university's online catalog, namely the study conducted by the Institute of Museum and Library Services of the University of Texas at Austin (2001). In one phase of their study, those researchers conducted focus groups with volunteers recruited from their university libraries staff; nearly three-fourths of the volunteers, being librarians from the public services cluster, were asked to represent "those library users who are served by the Web site and with whom the professional staff has contact on a regular basis" (Institute of Museum and Library Services [IMLS], 2000b). These librarians were asked to "think of a task that you typically do on UT Library on Line" and to "briefly describe this task" (IMLS, 2000a). In our study, we coded the list of their tasks to identify the most prevalent among them: finding details on a specific book, and finding materials on a specific topic, including searches of works by a given author.

From these categories we developed our tasks, which involved (1) identifying the most recent book in the library system written by a specific author, and (2) determining to which library or libraries a specific book belongs. These two tasks involve many of the same procedures as other tasks we did not test; they entail use of many of the same features of the site, and they require many the same skills on the part of the user. We believed, therefore, that these two tasks are representative of most if not all of the other tasks addressed by the online catalog, and so we operationally defined the catalog's success in terms of typical users' successful completion of these two tasks. We chose two tasks, so that they could be completed relatively quickly by subjects, e.g., in 10-15 minutes. This way we could test a relatively large number of users in order to do our *post hoc* retroactive analyses.

Usability Testing Procedure

Testing proceeded in the following manner. After identifying the campus buildings with the greatest number of computer laboratories available for student use, we visited the laboratories in their rank order, on different weekdays, and at various times of day. When the laboratories were crowded, we solicited participation of students waiting in line; otherwise, we asked them to participate at their workstation, working systematically through the laboratory. No more than eight subjects were recruited from any single laboratory, and no more than ten on any single day. Faculty and staff were selected in a similar manner. We identified the schools with the most students and visited the buildings on different days and at different times; we positioned ourselves at a haphazard location in the building and systematically solicited participation of faculty and staff at their desks. The community member was chosen by convenience and tested at home.

Two researchers from the team conducted each usability test, one facilitating the testing procedures, and the other recording observations regarding the subject's activities during the completion of the two designated tasks. In addition, the subjects completed a brief questionnaire of their computer experience and background information. Testing proceeded in this way until the target numbers in the sample strata were satisfied. The majority of the computer workstations featured Windows operating systems, though a small number of Macintosh machines were also used in the testing; all of the testing employed the Microsoft *Internet Explorer* software browser.

Analysis Procedure

We used a random number table to randomize the order of the usability test records, after the observations were completed. As mentioned above, we had collected data using from four to eight subjects from each computer cluster and had labeled the records chronologically. The purpose of randomizing the record order was to avoid possible bias relating to the data collection procedure. Data records were individually coded as either success or failure based on how well the subject had performed the tasks: specifically, if a subject succeeded on both tasks, this counted as a success; but if a subject failed both tasks, failed either of the two tasks, or only partially succeeded on one or both of the tasks, we coded it as a failure case. Next, we used the SPRT Web tool coded by Frick (2003) to analyze retroactively how many subjects would be needed to conclude whether the online catalog is effective or not. Finally, we changed various parameters of the SPRT, in order to see the effect number of subjects needed to reach a conclusion.

Results

We first defined the SPRT parameters as follows: If the online catalog Website were effective, we would expect 90% or more of the users to succeed in the tasks set to them. If the success rate were 60% or less, we would conclude that the site is not effective. We did not want to make false decisions more than five percent of the time, either way. Thus, we had:

Probability (success | Website is effective) = .90 or higher

Probability (success | Website not effective) = .60 or less

$\alpha = .05$

$\beta =. 05$

The first randomly selected subject from the pool of 51 subjects did not pass the test (this person failed on the second task). At this point we had observed one failure and no successes. The results are summarized in Table 2. The posterior probability for the site not working well became 0.80. The SPRT could not make a decision at this time.

The second randomly selected subject succeeded on both tasks, so altogether we had observed one success and one failure. The updated SPRT results are presented in Table 3. After this time, the posterior probability that the site was not working well dropped from .80 to approximately .73. Similar steps were repeated until the 12^{th} subject was tested. As it so happened, the remaining subjects were successful, so we had a total of 11 successes and the one initial failure. The SPRT results at this point are given Table 4. At this point in time, the posterior probability for site working well (0.956) had risen sufficiently to make a determination with the stopping rules. The probability ratio became 21.6, and so SPRT Rule 1 became true:

$$.956 / .0442 = 21.629 \geq (1 - \beta) / \alpha = 19$$

Accordingly, we stopped testing and concluded that the online catalog was working well. The inference is that no less than 90 percent of the target population would be successful in using the search engine for these two kinds of tasks. Our β error rate for reaching this conclusion falsely was 5 percent – we would be wrong in about 1 in 20 such studies, when the alternative is that the site would be ineffective if no more than 60 percent of users were successful. Most importantly, this is the same conclusion we would have reached if we had tested all 51 subjects. This demonstrates the value of sequential observations and the Wald decision rules.

We also noted that, given the same SPRT parameters had we not observed any failures in the first 8 subjects, we would have reached the same conclusion; conversely, if all of the initial subjects failed the tasks, SPRT would have resulted in a decision that the site was not working well with only 3 subjects.

In order to examine the behavior of the SPRT further, we analyzed the same randomly ordered data set under different conditions. Table 5 lists the results of running the SPRT while keeping alpha and beta errors constant (α=.05, β=.05) but changing the probabilities of success for effective and ineffective sites. The results reveal that as the zone of indifference shrinks, more subjects are needed to reach a decision regarding Website effectiveness. For example, if the success rate for a site not working well increases from 60% to 70%, the SPRT required an additional 6 successful users to reach the same decision. Notably, when comparing 90% vs. 80% for effectiveness versus non-effectiveness, no decision could be reached by the SPRT at the alpha and beta set at 0.05. Even though the overall success rate appears to be high (46/51 = 0.902), we still could not make a decision without the risk of concluding the site is effective, when in the target population it is not.

We also investigated the effect of keeping success and failure rates constant while reducing alpha and beta error levels. Table 6 indicates that as the alpha and beta error levels were reduced, more subjects were needed to reach a conclusion regarding Website effectiveness (e.g., to reduce the alpha and beta error levels from .05 to .01, the SPRT required an additional 4 successful users to reach a conclusion).

Finally, we set the expectancies for an effective web site to very high success rates compared with lower, but still high success rates for sites that would not be working well. As can be seen in Table 7, more observations were required to reach a conclusion compared with wider zones of indifference and lower rates in Table 5. And in each case, since our expectations for the maximum success rate for a Website that was not working well was quite high, the conclusion reached in all cases was that the site was not effective.

Discussion

In this study of Website effectiveness, the usability test results (in which subjects were subsequently scrambled into a random order) were analyzed retroactively by the SPRT to determine whether the site was successful or not. The study provides evidence of the potential of the SPRT in usability testing where determination of effectiveness rather than error detection is the goal. The SPRT affords a simple and sound alternative to raw percentages or statistical procedures such as the Bayesian BETA distribution, which are likely to require more users at the same error rates (cf., Frick, 1990).

At first glance, the requirement of *a priori* specification of P_e and P_n may seem to be a limitation of SPRT. Rather than testing discrete alternatives as does the SPRT, one could compare the alternatives that the success rate is 0.85 or higher versus being less than 0.85. The Bayesian

method for doing this utilizes the BETA distribution (Schmitt, 1969). However, as Frick (1990) demonstrated in Monte Carlo simulations, the sequential decision procedure with this Bayesian method is more likely to result in α errors early on in the sampling (choose alternative II | alternative I really true), compared to the SPRT, and the posterior BETA distribution often required larger numbers of subjects to of choose alternative I with the same β error rate. We could use the posterior BETA distribution to estimate a 90 percent confidence interval for the success rate, but when doing so the decision maker is still faced with the issue of how narrow that confidence interval should be in order to be satisfied. For example, in the case of 11 successes and 1 failure, the observed success rate is 11/12 = 0.92. The 0.90 highest density region of the BETA distribution with a flat prior distribution is 0.707 to 0.990 (Schmitt, 1969, p. 379). Thus, we would be 90 percent confident that in the target population the success rate would be somewhere between 71 percent and 99 percent, given our empirical observation of 11 successes and 1 failure. Choosing the 90 percent confidence interval would leave 5 percent on each tail of the distribution, which would be equivalent to setting $\alpha = 0.05$ and $\beta = 0.05$.

Alternatively, one can use estimation procedures outlined by Ferguson (1971) or Hays (1973) for using the normal distribution to estimate a confidence interval. However, as noted earlier in this article, the Gaussian distribution is problematic for such estimation when success rates are very low or very high or when the sample size is relatively small.

In general our conclusions are consistent with those in Monte Carlo studies conducted by Frick (1990) when comparing SPRT with BETA and item response theory models for making decisions. The same patterns of results occurred, even though that study was concerned with computer-adaptive mastery testing, whereas this study was concerned with how many subjects are need in a usability test to determine Website effectiveness. All other things being equal, when the

zone of indifference becomes narrower, larger sample sizes are needed; and all other things being equal, when lower the error rates for α and β are specified, larger sample sizes needed. These patterns can be confirmed as well by using Frick's (2003) Web tool.

What the SPRT requires the decision maker to do is specify *a priori* acceptable success rates for effective and ineffective Websites, respectively, and to specify the likelihoods of making α and β errors in drawing conclusions about the site's effectiveness based on the number of subjects tested. Decision makers may not be used to such requirements. They may not accustomed to statistical decision making – i.e., making inferences from a sample to a target population – and they may arrive at erroneous conclusions about the effectiveness of their Websites without realizing it, by not taking into account sampling error. Ignorance may be bliss, but when it really matters if a Website is indeed working well or not, then the SPRT (or other inferential statistical procedures) will make explicit how big a gamble is really being taken. For that matter, if probability theory were well understood, people would not go to gambling casinos and throw away their own money – unless they find that entertaining and enjoyable. The SPRT requires that the decision maker be relatively precise in stating the nature of the gamble he or she is making when betting on whether the Website will work well or not with the target population.

The SPRT, when used appropriately, also requires that subjects be selected in a random manner. Convenience samples may be just that – relatively easy to get – but may not be representative of the target audience. If the sample is unrepresentative, then the statistical inference is likely to be invalid. As stated earlier, if a relatively large pool of representative users is selected in advance, and if one selects randomly from this pool one user at a time for a usability test and conducts the SPRT after each usability test, then this is a practical way to help improve the validity of the decision reached. This may not be perfect – i.e., strictly satisfying the requirements

for random selection – but it appears to be a reasonable compromise to keep usability testing expenses from getting out of hand. Perhaps someday we will have enough trained usability testing specialists located in many places who can help carry out such tests with subjects selected in a more pure random manner. Or perhaps we will be able to observe users carry out authentic tasks with Websites over a distance, for example, by two-way videoconferencing.

While not bearing upon the usefulness of SPRT in usability testing, a few points regarding actual testing in our study also deserve to be mentioned. First, the percentage of failures encountered during the study needs qualification. In several cases, despite the subject's entry of the correct information using the correct submission procedures (e.g., conducting a "title search" of "all libraries"), the server produced incorrect results, that is, results inconsistent with the results produced under the same conditions at other times; despite the fact that the subject used the online catalog in the correct manner, we tallied this as evidence against the site's effectiveness. Further, in most of the testing situations, the subjects experienced inordinate server delays in receiving results; many subjects interpreted this as an error on their part and returned to the search page to review their input, or repeatedly clicked the submit buttons, or in other ways disrupted the original usage scenario. In every case, we let the encounter proceed to its conclusion—often to success, however slow, but in several cases converting what would have been a successful case to one of failure to accomplish the task. Not only did the server, through its errors and delays, contribute to the number of unsuccessful searches, but our own criteria for success may be regarded as unduly stringent. Specifically, only if a subject succeeded on both of the tasks did we regard the case as a success; if the subject was successful on one task but only partially successful on another, we counted the entire case as a failure – a definition of success perhaps not reflective of the Website owner's own, but one that ultimately provided data suitable to SPRT analysis.

A second consideration was the inconsistency of the appearance of the search page in different situations. Specifically, the HTML coding of the search page specified that, in the drop-down list from which the user selects which libraries to include in the search, the default or selected option is “all campus libraries,” meaning all libraries on the local campus but excluding all libraries on other campuses. In common settings, this default setting is used to guide the search, unless the user selects otherwise, but in the computer laboratories available for student use, this default is overridden: the browser instead presents “all libraries,” that is all libraries on all campuses, as the default. This variation resulted in inconsistencies among results. Since the participants were solicited by convenience, their investment in the testing was likely only casual, and indeed, while the tasks were commonplace, they were not intrinsic. Consequently, though both of the tasks called for the subject to find a reference from any of the libraries within the university system, one of the tasks addressed an item located only in an off-campus library. On this task, then, users at computers other than the campus laboratory workstations would have had to change the option relating to library selection to retrieve the same results as users in the laboratories, that is, to find the correct reference; otherwise, a different result would consistently be returned by the search engine. We considered this difference to be a limitation of the testing procedures (e.g., subjects recruited without compensation) rather than a limitation of the Website (though the default option bears significant implications on the usability of the system), and accordingly, for users whose default setting covered only campus libraries, we accepted the alternate answer as correct. As with the errors discussed above, this limitation may bear upon accepting the findings as representative of the site’s usability, but not upon the usefulness of SPRT procedures in usability testing more generally.

Finally, a related consideration is the limitation of generalizing from the usability tasks to the catalog search engine more broadly. While several of the features were not tested directly (e.g., searches for journal titles), we nonetheless consider them to be similar in presentation and functionally to the tasks covered by the testing. Accordingly, we may tentatively generalize the site's effectiveness on the tasks tested to reflect the site's effectiveness for the related tasks. Still, this step is problematized by the interaction between the tasks and the libraries searched, but again, this does not pertain to the SPRT analysis.

While the study offers data regarding the usability of a particular Website search engine, and while the methods and usability results may inform future studies of Website effectiveness, the chief contribution of this study is the demonstration of SPRT's application in usability testing. Further studies may likewise contribute to this body of knowledge through several avenues of inquiry: they may continue comparing SPRT to other statistical procedures to establish its benefits and limitations; explore the range of applications of SPRT to gauge its usefulness and flexibility; and establish methods of implementing SPRT during testing to determine when to stop testing. We hope that the present study demonstrates the promise of such pursuits.

Acknowledgements

We want to thank Dr. Todd Zazelenchuk, now a usability specialist at *Whirlpool Corporation,* for his contributions to the section on usability definitions. We also thank Dr. Joanne Peng, a professor at Indiana University and expert in inferential statistics and experimental research design, for critically reviewing the sections on statistical inference with the SPRT and decision error rates.

References

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

Colton, T., & McPherson, K. (1976). Two-stage plans compared with fixed-sampling-size and Wald SPRT plans. *Journal of the American Statistical Association, 71.*

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

Ferguson, G. A. (1971). *Statistical analysis in psychology and education* (third edition). New York: McGraw-Hill.

Fisher, R. A. *Statistical methods and scientific inference*. New York: Hafner, 1956.

Frick, T. W. (1989). Bayesian adaptation during computer-based tests and computer-guided practice exercises. *Journal of Educational Computing Research, 5*(1), 89-114.

Frick, T. W. (1990). A Comparison of Three Decision Models for Adapting the Length of Computer-Based Mastery Tests. *Journal of Educational Computing Research, 6*(4), 469-503.

Frick, T.W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8*(2), 187-213.

Frick, T. W. (2003). Web Tool for Sequential Bayesian Decision Making. Retrieved April 29, 2004, from http://education.indiana.edu/~frick/decide/start.html.

Hays, W. L. (1973). Statistics for the social sciences. (Second Edition) New York: Holt, Rinehart and Winston.

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*(4), 421-443.

Hudson, W. (2001). How many users does it take to change a web site? *SIGCHI Bulletin*, May/June 2001. Retrieved October 8, 2002, from http://www.syntagm.co.uk/design/articles/howmany.htm

Institute of Museum and Library Services (2000a). Evaluation of Effective Library Sites. *Characterization of the use and usability of a web-based digital library* (Usability Testing section). Retrieved October 6, 2002, from http://imls.lib.utexas.edu/usability/evalmit.html.

Institute of Museum and Library Services (2000b). Report Covering Project Activities to 6/30/2000. *Characterization of the use and usability of a web-based digital library* (Reports section). Retrieved October 6, 2002, from http://imls.lib.utexas.edu/report/activities00-06-30.html.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The evaluator effect in usability tests. In C.-M. Karat & A. Lund (Eds.), *Human Factors in Computing Systems CHI'98 Summary* (pp. 255-256). New York: ACM Press.

Kirk, R. E. (1995). *Experimental design: Procedures for the social sciences*. Pacific Grove, CA: Brooks/Cole.

Kirkpatrick, D.L. (1994). *Evaluating training programs: The four levels*. San Francisco, CA: Berrett-Koehler.

Krug, S. (2000). *Don't make me think: A common sense approach to Web usability*. Indianapolis, IN: New Riders.

Lewis, C., & Sheenan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 376-386.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors*, *36*(2), 368-378.

Lewis, J. R. (2001). Introduction: Current issues in usability testing. *International Journal of Human-Computer Interaction*, *13*(4), 343-349.

Littlewood, B., & Wright, D. (1997). Some conservative stopping rules for the operational testing of safety-critical software. *IEEE Transactions on Software Engineering, 23*(11).

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D. et al. (1998). Comparative evaluation of usability tests. In *Proceedings of the Usability Professionals Association 1998 (UPA98) Conference* (pp. 189-200). Washington D.C.: Usability Professionals Association. Retrieved October 8, 2002, from http://www.dialogdesign.dk/tekster/cue1/cue1paper.doc

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

Nielsen, J. (2000a). Why you only need to test with 5 users. *Alertbox*, March 19, 2000. Retrieved October 8, 2002, from http://www.useit.com/alertbox/20000319.html.

Nielsen, J. (2000b). *Designing web usability*. Indianapolis, IN: New Riders.

Nielsen, J. (2001). Success rate: The simplest usability metric. *Alertbox*, February 18, 2001. Retrieved November 13, 2002, from http://www.useit.com/alertbox/20010218.html

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of INTERCHI '93* (pp. 206-213). Amsterdam, The Netherlands: ACM Press.

Preece, J., Ed. (1993). *A guide to usability: Human factors in computing*. Wokingham, Addison-Wesley.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–255). New York: Academic Press.

Seels, B.B. & Richey, R.C. (1994). *Instructional Technology: The definition and domains of the field*. Washington, C.C.: Association for Educational Communications and Technology.

Schmitt, S. A. (1969). *Measuring uncertainty: An elementary introduction to Bayesian statistics*. Reading, MA: Addison Wesley Publishing Company, Inc.

Shackel, B. (1986). People and computers: Designing for usability. Conference of the British Computer Society: Human Computer Interaction Specialist Group, University of York, UK, Cambridge University Press.

Shackel, B. (1991). Usability - context, framework, definition, design and evaluation. In B. Shackel and S. Richardson, *Human Factors for Informatics Usability*. Cambridge, England: Cambridge University Press: 21-38.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (third edition). Reading, MA: Addison-Wesley.

Spool, J., & Schroeder, W. (2001). Testing web sites: Five users is no where near enough. In J. Jacko & A. Sears (Eds), *Conference on Human Factors in Computing Systems: CHI 2001 Extended Abstracts* (pp. 285-286). Seattle, WA: ACM Press.

Turner, C. W., Nielsen, J., & Lewis, J. R. (2002). Current issues in the determination of usability test sample size: How many users is enough? In *Usability Professionals' Association 2002 Conference Proceedings* (n.p.). Chicago: UPA.

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. *Human Factors and Ergonomics Society 34th Annual Meeting* (pp. 291-294). Santa Monica, CA: Human Factors and Ergonomics Society.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, *34*(4), 457-468.

Wald, A. (1945). Sequential method for deciding between two courses of action. *Journal of the American Statistical Association, 40*(231), 277-306.

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). Toulouse, France: Cépadèus Éditions. Retrieved October 8, 2002, from http://www.cet.sunderland.ac.uk/~cs0gco/fiveusers.doc

Table 1. *Three popular definitions of usability (van Welie, van der Veer et al. 1999)*

ISO 9241-11	**Nielsen (1993)**	**Shneiderman (1998)**
Efficiency	Efficiency Learnability	Speed of Performance Time to Learn
Effectiveness	Memorability Errors/Safety	Retention over Time Rate of Errors by Users
Satisfaction	Satisfaction	Subjective Satisfaction

Figure 1. Illustration of the zone of indifference for SPRT decision making.

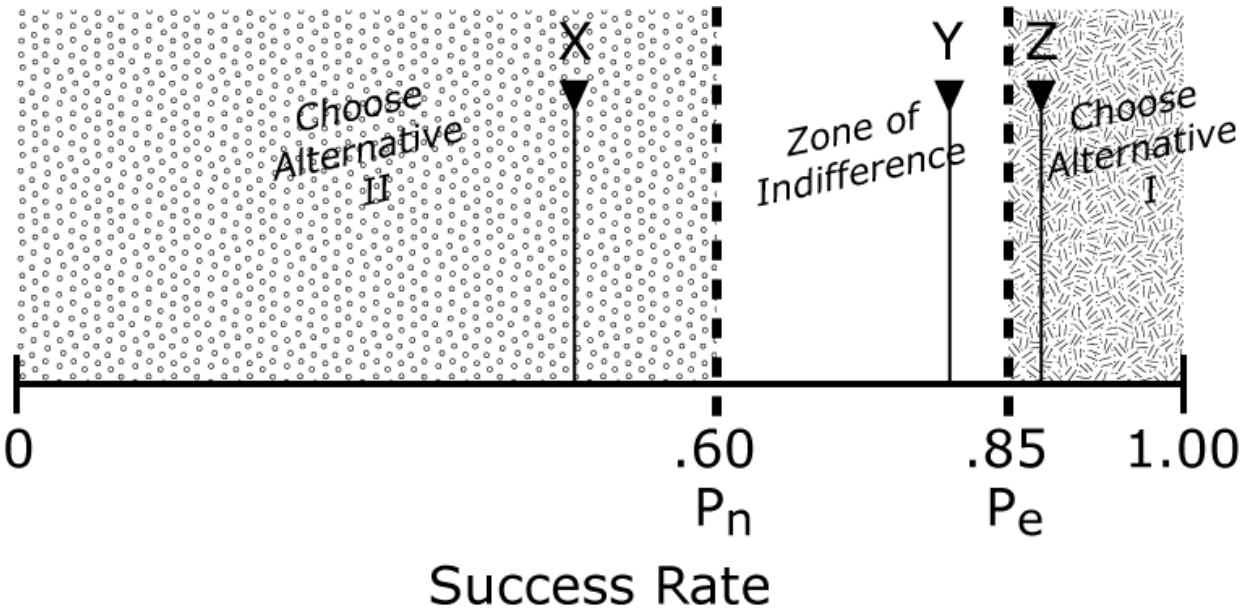

Figure 2. Illustration of the online card catalog search interface.

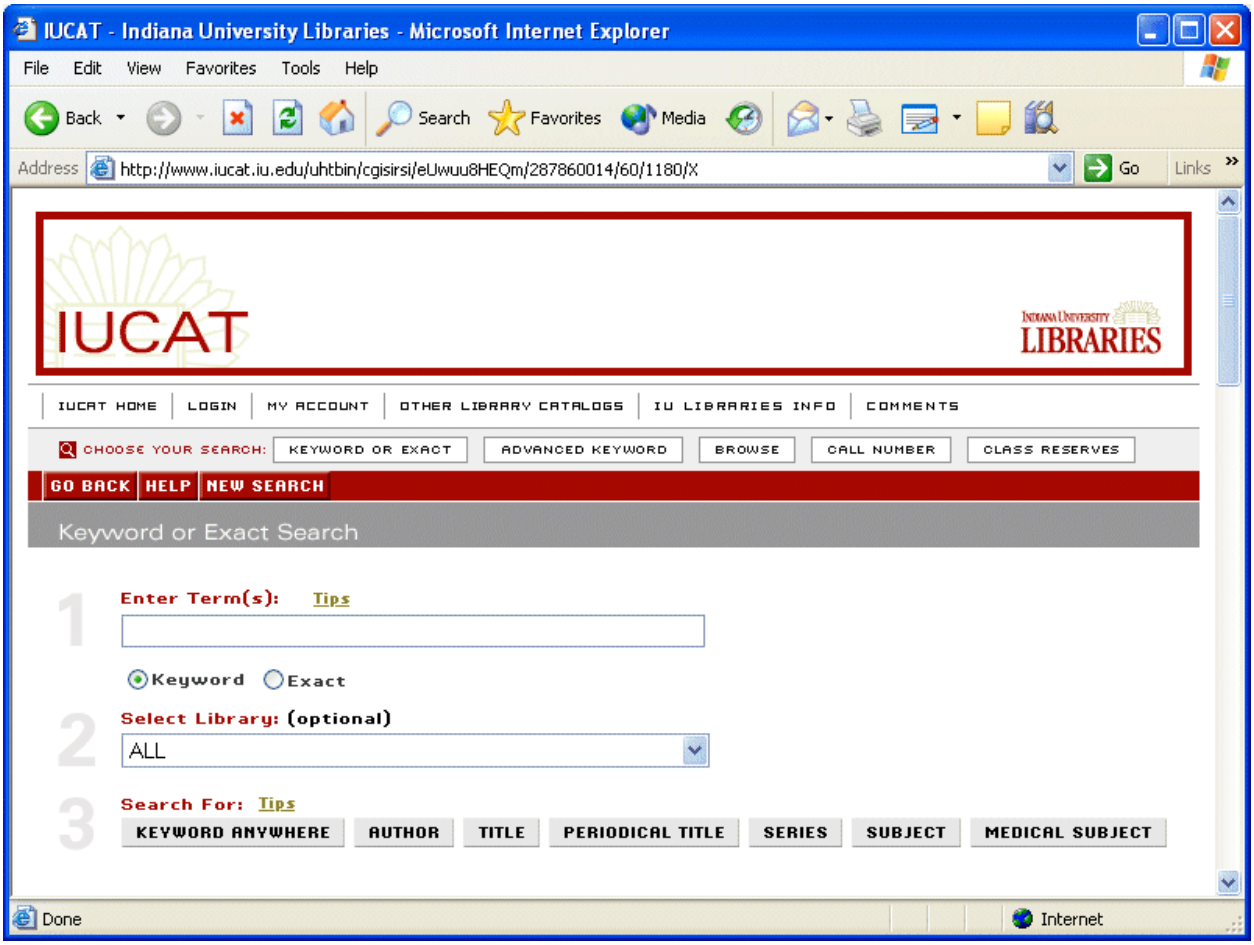

Table 2

SPRT Results after 1 Subject who Failed the Usability Test

	Probability			
Alternative	Prior	Conditional	Joint	Posterior
Site working well	.5000 ×	.1000 =	.0500 / sum =	.2000
Site NOT working well	.5000 ×	.4000 =	.2000 / sum =	.8000
			sum = 0.2500	

Table 3

SPRT Results after 2 Subjects: the Second Subject Succeeded in the Usability Test

Alternative	Probability			
	Prior	Conditional	Joint	Posterior
Site working well	.2000 ×	.9000 =	.1800 / sum =	.2727
Site NOT working well	.8000 ×	.6000 =	.4800 / sum =	.7273
			sum =.6600	

Table 4

SPRT Results after 12 Subjects: A Total of 11 Successful Subjects and One who Failed.

	Probability			
Alternative	Prior	Conditional	Joint	Posterior
Site working well	.9351 ×	.9000 =	.8416 / sum =	.9558
Site NOT working well	.0648 ×	.6000 =	.0389 / sum =	.0442
			sum = 0.8805	

Table 5

SPRT Results with α=.05, β=.05

Success Rate		Observations			
Working well	Not working well	Successes	Failures	Total Users	Conclusion
90%	50%	8	1	9	Effective
90%	60%	11	1	12	Effective
90%	70%	17	1	18	Effective
90%	80%	46	5	51	No conclusion

Table 6

SPRT Results with Success Rate for Working Well = 90%, for Not Working Well =60%

Level		Observations			
α	β	Successes	Failures	Total Users	Conclusion
0.05	0.05	11	1	12	Effective
0.03	0.03	12	1	13	Effective
0.01	0.01	15	1	16	Effective

Table 7

SPRT Results with α=.05, β=.05

Success Rate		Observations			
Working well	Not working well	Successes	Failures	Total Users	Conclusion
98%	90%	31	4	35	Not effective
99%	90%	24	3	27	Not effective
99%	98%	41	5	46	Not effective