

# **Extreme Discount Usability Engineering**

**Paul F. Marty**

College of Information  
Florida State University  
Tallahassee FL 32306-2100

**Michael B. Twidale**

Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
Champaign, IL 61820

## **Abstract**

This paper explores the circumstances under which extremely discounted usability engineering techniques produce results that are worthwhile to researchers and practitioners. We present a method for usability analysis where an entire evaluation can be conducted in only thirty minutes, an amount of time we suggest as perhaps the smallest possible unit of analysis (or quantum) for complete usability evaluations. Based on a quantitative and qualitative analysis of thirty-six separate trials of this method, we assess the value of discount usability engineering at extremes of time and tests, and present our observations about the nature of user testing when pushed to extreme limits, a phenomenon we call the “Quantum Usability Effect.”

## **Keywords**

Usability Analysis; User Testing; Discount Usability Engineering; Research Methods; Experimental Design;

# Extreme Discount Usability Engineering

## 1. Introduction

Over the past few years, there have been many controversies over the value of discount usability engineering techniques for evaluating the usability of information interfaces (Wixon, 2003). In most of these arguments, the established, scientific value of formal user testing is pitted against the efficiency and cost-effectiveness of less formal usability analysis methods (cf. Nielsen and Mack, 1995). Advocates of the discount usability engineering approach argue that a small number of relatively informal user studies can discover the majority of usability problems, and that these results can be used to improve the usability of information interfaces, even if they lack the scientific rigor of more formal methodologies (Nielsen, 1994a).

Resolving this argument is important, since traditional user testing is still often seen as time-consuming and expensive, and usability advocates worry that unless an alternative to more formal methods is available, many organizations will continue to forgo user testing completely (Dumas, 2002). Researchers interested in discount usability engineering, therefore, focus their efforts on minimizing the cost and time required for user testing and maximizing the benefits accrued from those tests (Bias and Mayhew, 1994). These researchers have conducted many studies to answer questions about the relative effectiveness of various usability evaluation methods or the minimum number of user tests that will result in the maximize return on investment for usability evaluators (Jeffries et al., 1991; Desurvire, 1994). Nevertheless, many doubt whether discount usability engineering techniques produce results that are worthwhile or even scientifically valid (Gray and Saltzman, 1998).

Recently, the authors have had the opportunity to study this issue in some detail. Over the past three years, we have been performing a study that has enabled us to explore a lower bound on the value of discount usability engineering. While conducting a series of high-speed, low-cost usability demonstrations for information professionals, we began to wonder what would happen if the idea behind discount usability engineering was taken to extremes, in terms of the amount of time spent on each test and the number of user tests conducted. If usability evaluators employed analysis techniques that were as fast and as cheap as possible, would they still be better than nothing? Exactly how discounted can discount usability engineering become, we wondered, before it no longer produces worthwhile results?

To discover the answers to these questions, we devised a method for usability analysis where an entire evaluation could be conducted in only thirty minutes, an amount of time we suggest as perhaps the smallest possible unit of analysis (or quantum) for complete usability evaluations. We divided each thirty-minute evaluation into ten minutes for task analysis, ten minutes for user testing, and ten minutes for analysis of results. Given that most usability evaluations take a minimum of fifty hours to complete the same process (Nielsen, 1994a), with at least an hour dedicated to each individual user test on average, we wondered whether these very brief evaluations would allow us to uncover any usability flaws valuable to system designers. To our surprise we found that we were consistently able to transition from knowing absolutely nothing about the usability of an interface to making valuable recommendations for improving its design in thirty-six of thirty-six separate trials of this method.

This article, therefore, will analyze the results of our thirty-six evaluations, assessing the value of discount usability engineering at extremes of time and tests. We will demonstrate that trained usability evaluators can study a previously unfamiliar interface, assess its strengths and weaknesses, develop representative scenarios of use, administer these scenarios to representative user testers, analyze and evaluate the results, and generate relevant and useful recommendations for design improvements in thirty minutes with one user test. We will also present our observations about the nature of user testing when pushed to extreme limits, a phenomenon we call the “Quantum Usability Effect,” along with its implications for the future of research into discount usability engineering.

## **2. Discount Usability Engineering: Faster, Cheaper, and Better?**

An aphorism famous in the engineering community states: “better, faster, cheaper; choose two.” The challenge of making products or processes that are at once “good, fast, and cheap” underlies the problems faced by advocates of discount usability engineering methods (Sauro, 2004). This form of usability analysis represents an attempt to reduce the cost and time necessary to evaluate interfaces for usability while maintaining the high value of user testing for improving interfaces to information systems (Nielsen, 1994a). To accomplish this goal, advocates of the discount usability engineering approach need to demonstrate that usability evaluation methods which are cheap and fast can also be good; therefore, they have focused on providing cost-justifications for usability analysis, resulting in astounding figures for a return on investment for usability analysis (Bias and Mayhew, 1994; Donahue, 2001; Siegel, 2003).

This emphasis on demonstrating the cost-benefits of usability analysis has prompted researchers and practitioners to search for the best way of maximizing return on investment, resulting in numerous studies that evaluate the relative effectiveness of various usability methods (Jeffries et al., 1991; Desurvire, 1994; Karat et al., 1992; Nielsen and Phillips, 1993). Determining the most appropriate criteria for comparing usability evaluation methods, however, is a question that has proven to be extremely controversial (Gray and Saltzman, 1998). Many researchers argue that the best way of determining what method of usability analysis provides the most return on investment is to compare the number of usability flaws discovered by a particular method with the amount of effort expended in implementing that method (Wixon, 2003). This approach raises certain questions.

This most controversial question faced by advocates of discount usability engineering is: how much user testing provides the most return on investment? Many studies have explored the question of how many users need to be tested before evaluators will learn the majority of what they need to know about the usability of any given interface (Virzi, 1992). One popular theory claims that the law of diminishing returns kicks in after approximately five user tests (Landauer and Nielsen, 1993; Nielsen, 1994b). Many people have tried to either support or disprove this theory, with most recent research focused on the question of whether five users is enough, and many researchers arguing that more user tests are necessary (Spool and Schroeder, 2001; Faulkner, 2003; Woolrych and Cockton, 2001). What is interesting, however, is that virtually no one is arguing that usability evaluators should focus on fewer than five users.

Most discount usability researchers remain focused on the “high end” of discount usability engineering; they are generally attempting to define acceptable upper bounds rather than trying to establish acceptable lower bounds. They are determined to discover the smallest amount of effort (number of user tests) that will return the greatest amount of data (number of usability flaws), thereby proving that discount methods can be fast and cheap and also good. This mentality keeps them focused on the “top-down” question of how much to discount traditional user testing methods while maintaining their overall effectiveness, rather than the “bottom-up” question of how little effort evaluators need to expend before a method returns anything at all.

What would happen, we wondered, if advocates of discount usability engineering stopped worrying about whether their methods were good, but simply asked whether their methods were better than nothing? What if usability evaluators stopped trying to find the majority of usability problems, were no longer concerned with being thorough, but were interested in seeing how quickly they could learn anything at all about an interface that might be useful to designers? One immediate result of this approach is that it would take us away from the notion that the best way to assess usability evaluation methods is to count the number of usability flaws discovered.

When researchers focus on determining the number of usability problems detected per user test, they tend to look at these results in isolation from their ability to improve user interfaces. According to Wixon (2003), comparing usability evaluation methods based solely on the number of problems found per user tested has distracted us from the main problem:

“Consider the debate about how many participants are needed. It is focused exclusively on problem detection—that is, how likely that we will find a given subset of the problems. In a practical world, applying such criteria to evaluating methods [...] ignores that problems should be fixed and not just found. If we considered instead a more relevant criterion—namely, how much can we improve the product in the shortest time with the least effort?—we probably would not have asked this question...” (Wixon, 2003, p.31)

Is it possible that usability evaluation methods that seek to uncover as many usability flaws as possible could be less important than methods that attempt to find something constructive that designers can use to improve an interface, even if those results are not 100% scientifically valid? Nielsen (1994a) touches upon this idea when he writes that he will “focus on achieving ‘the good’ with respect to having some usability engineering work performed, even though the methods needed to achieve this result are definitely not ‘the best’ method and will not give perfect results” (p. 249). Usability evaluators have also explored this question in industry (Jordan et al., 1996), where time is of the essence, and results from “quick and dirty” usability evaluations are more important than following formal user testing methods (Thomas, 1996). It does not seem, however, that this approach has been embraced by many researchers who pursue discount usability engineering.

The worry, of course, is that evaluators who attempt to “improve the product in the shortest time with the least effort” without first performing formal usability evaluations risk results that are completely worthless. Discount usability engineering methods need to be fast and cheap, but not so fast and so cheap that they are no longer valid. As Thomas (1996) points out, the important

thing is to find out “when does quick and dirty become fast and filthy?” (p. 113). This is a good question, but the trick to answering this question lies in understanding that new methods of usability analysis require new approaches to evaluating those methods. In order to know when too fast and too cheap is no longer better than nothing, usability researchers need a better understanding of what happens at the extremes of user testing—minimal amounts of time, minimum number of users—what we might call the “quantum effect” of usability analysis.

Few usability researchers have spent much time exploring these extremes of usability analysis. Much as in the world of particle physics, very strange things happen when evaluators deal with very small amounts of usability analysis: the “traditional laws” of user testing, for one, appear to no longer hold true. Working at this level of analysis is very difficult, validating results can be next to impossible, and yet usability evaluators need to know what happens as they proceed from running no tests to running one test, from spending no time with an interface to spending a little time with an interface, from knowing nothing to knowing something. Researchers and practitioners interested in the “quantum effect” of usability analysis face two challenges.

The first challenge lies in the difficulty of spending as little time as possible conducting user tests. Nielsen (1994a) estimates that the process of conducting one user test (including evaluating the interface, planning the test, coming up with representative tasks, administering those tasks, evaluating the results, and making design recommendations) can consume a minimum of 50 hours. Every stage in user testing is time consuming, and for many usability engineers, high-speed user testing means limiting the testing process to three days (Bauersfield and Halgren, 1996; cf. Wixon and Ramey, 1996). These numbers lead most evaluators to suspect that one user test conducted in mere minutes would not be a constructive method of evaluating an interface. Researchers interested in studying how usability analysis can be conducted in extremely small increments of time will need new methods of producing useful recommendations for design as quickly as possible as well as new ways of evaluating the validity of those recommendations without spending any additional amount of time in the process.

The second challenge lies in the difficulties inherent in conducting only one user test. From a research perspective, there are many serious problems with this approach, including risks to the completeness, coverage, and validity of results. On the other hand, most evaluators agree that testing once is better than not testing at all, and some even claim that one user test can return about 25% to 40% of all usability problems with an interface (Nielsen and Landauer, 1993). As Dicks (2002) writes, “even though the results of using one of the “discount” usability methods or of using small samples may not stand up to the rigors of controlled experimental enquiry, they can still yield very useful benefits to practitioners” (p. 30). While the validity of this approach can be easily dismissed by individuals conducting multiple tests to determine all possible problems with an interface, researchers exploring the effectiveness of user testing at these extremes must find a way to maximize the ability of one test to yield useful results. If usability evaluators are to learn useful things from conducting one user test, they need new methods of assessing usability and evaluating their results.

In order to cope with both of these challenges, therefore, researchers working at the quantum level of usability analysis require a new approach to usability assessment, one that does not depend on comparing the results from one user test to results from multiple tests, but on

exploring how quickly usability evaluators transition from knowing absolutely nothing about an interface to making constructive recommendations for design improvements. This mentality is very different from the typical concerns of usability evaluators, but it can be a very important one in situations where time and money put serious constraints on traditional user testing methods. A study of usability techniques capable of dealing with these extremes of tests and time will help us establish a lower bound on the ability of discount usability engineering to produce worthwhile results.

### **3. Developing and Assessing a Method for User Testing at “Quantum Extremes”**

In order to better our understanding of the nature of user testing at extremes of time and tests, we developed a method of high-speed, low-cost user testing that would enable us to conduct a complete usability evaluation in thirty minutes with one user test. Our original purpose in developing this method was to advocate for increased usability analysis of museum websites at national and international conferences for museum professionals (Marty and Twidale, 2004). To do so, we needed a method that would allow us to illustrate quickly the power of user testing and the need for frequent usability evaluations in the museum website development process. As these demonstrations proceeded, however, we realized that we had at our disposal a method that would allow us to compare the value of discount usability engineering at “quantum extremes.”

While a detailed step by step account of the workings of this method is beyond the scope of this article, the method requires usability evaluators to spend ten minutes assessing a previously unknown interface and developing representative tasks, ten minutes administering these tasks to representative users, and ten minutes analyzing the results of the tests to identify usability flaws and make recommendations for design (cf. Marty and Twidale, manuscript under review). This method is different from that of traditional user testing, where a single website is tested at length, with many subjects, in order to validate the overall usability of the site (in the case of a summative evaluation) or uncover its major and minor usability flaws (in the case of a formative evaluation). The emphasis of our method is not on creating an exhaustive list of all, or even most, of the usability flaws in any given interface, but on quickly finding a non-zero number of usability flaws that the participants involved in the method believe will improve the design of the interface.

To assess the capabilities of this method, we conducted thirty-six separate trials of the method at five different national and international conferences for museum professionals over the past three years. The websites we analyzed were all museum websites suggested by audience members. The attendees at these conferences were typically users of museum websites, developers of museum websites, or both. A representative from the organization whose website was being evaluated was required to be on hand during the user testing session. The user testers were volunteers selected from the audience at the start of each test; they had no prior familiarity with the specific museum website being evaluated. In order to learn what the volunteer user testers were thinking, we employed two common techniques (usually alternating them from test to test). In 20 of 36 trials, we used the “think-aloud” protocol where one volunteer user vocalized his or her thoughts during the test (Waes, 2000). In 16 of 36 trials, we used a variation on “constructive interaction” where two volunteers discussed strategies of completing tasks in front of the audience (Wildman, 1995).

Throughout the process, the authors (who also served as the usability evaluators) took notes about each test. Conducting research at such high speeds is not easy, and the public nature of these events combined with the impossibility of obtaining informed consent from all audience members attending these conferences meant that it was not possible for us to create audio or video recordings of the tests. We therefore had to rely on notes that allowed us to reconstruct each test at the end of each conference; while necessarily brief, these provided valuable data that helped us assess the method. We recorded the types of sites evaluated, the tasks developed for each test, the number of tasks actually administered during the test, the number of tasks completed by the user testers, the usability flaws uncovered in each test, and design recommendations made by the evaluators. We also gathered informal feedback from the participants in each demonstration by asking the site representatives, the volunteer user testers, and the audience whether they thought that the results of each test represented valid usability problems that typical users might face.

It is important to understand that these thirty-six trials were not the same experiment replicated thirty-six times, but an activity pattern uniquely instantiated with thirty-six different websites. In each evaluation, different tasks were set, and different events occurred, all tailored to the particular needs and characteristics of each website. In this study, we are not comparing results from site to site or from test to test, but checking the validity of the method against itself, in thirty-six different trials. We therefore conducted a quantitative and qualitative analysis to determine the ability of the usability evaluators to develop and administer tasks in short periods of time, the ability of user testers to complete tasks under these constraints, and the overall number of usability flaws and design recommendations that can be determined in a very short amount of time. By this analysis, we were able to explore the nature of user testing at extremes of time and tests, compare our approach to traditional methods of usability analysis, and develop observations about user testing and the “quantum usability effect.”

#### **4. Looking for Results from Thirty-Six Extremely-Discounted User Tests**

Our data analysis illustrates that our method was extremely successful in quickly finding a non-zero number of usability flaws with the interfaces we evaluated. In thirty-six thirty-minute usability evaluations (for a total of 18 hours), we developed a total of 151 representative tasks and administered 119 of them to representative users, who were then able to complete 83 of those tasks during the tests. Moreover, the results of these tests enabled us to make more than 500 design recommendations, each validated by the method participants (see below), and each directly related to at least one usability flaw found while the volunteers were attempting to complete the administered tasks (see Table 1).

During each thirty minute evaluation, we developed between one and eight tasks (at a rate of approximately one per minute); we administered between one and five tasks (usually administering three of every four tasks developed); and volunteers user testers completed between zero and five tasks (usually succeeding in completing two out of three tasks). By the end of each evaluation, moreover, we had made somewhere between five and twenty-five design recommendations to improve the website, each reflecting usability flaws uncovered by our analysis. On average, therefore, in one thirty-minute test, we were able to develop approximately

four tasks; we were able to administer approximately three tasks; the users could successfully complete approximately two tasks; and we could make somewhere between ten and fifteen recommendations for design (see Table 2).

Most notable about these results was that there was never an instance of zero findings in any of the thirty-six evaluations (see Table 3). At a minimum, we always found at least five to ten usability flaws which we were able to translate into recommendations for design improvements. At no time did we have nothing to say, nor were these stock responses we kept in reserve from some kind of prior evaluation; there were no prior evaluations conducted of any of the sites. All of these findings were specific recommendations that derived directly from the results of the tests.

Given the nature of the method, we cannot claim, of course, that these findings were in any way comprehensive or indicative of the majority of usability problems with the evaluated sites. It is also likely that the flaws we caught this way were the really big ones, the “obvious” ones that must be fixed first. While it is not too surprising that the low-hanging fruit are easier to catch, what is surprising is just how low the low-hanging fruit hangs, given the right methods. The flaws we found using this method were discovered almost immediately, and therefore these numbers indicate that evaluators can develop valuable recommendations for design, even in non-ideal situations and very short periods of time.

Also, we cannot claim that these findings will hold true over multiple tests; having only one user test in our evaluation, we could not use formal, traditional methods to validate our findings. In each test, however, we informally validated the results of our test by noting how particular findings were often observed in other websites and widely reported upon in the literature on website usability. In addition, we informally validated our findings with the site representatives, volunteer user testers, and audience members by checking to see if our findings would be taken seriously by observers and participants. We found that, in every case, the participants recognized the usability flaws we discussed as serious usability problems and agreed that implementing the design recommendations would in fact improve the usability of the site being evaluated.

What we can say is that in thirty-six of thirty-six evaluations, we consistently produced non-zero recommendations for design, and all participants (evaluators, volunteer user testers, site representatives, and audience members) confirmed the validity of the findings in terms of usability problems that needed to be solved. Whether or not this method produced either the majority or the most significant usability problems with the sites being tested is irrelevant. All that mattered was that each participant agreed that the findings were relevant, the evaluations were valuable, and the recommendations for design would result in overall usability improvements. In the end result, whether our design recommendations would actually make the site better was not as important as convincing the participants that through discount usability engineering methods, they could make their site better. Whether our recommendations for design would be considered objectively valid by outside observers or would remain consistent over multiple tests was not as important as the fact that we consistently found something, and that what we found was considered valid by the participants in the method.

These results indicate that when conducting and evaluating user tests at such “quantum extremes,” it is necessary that usability evaluators have a new way of thinking about the overall purpose of user testing. If interface designers are to benefit from extremely-discounted usability analysis methods such as the one used in this study, it is important that researchers and practitioners understand the implications of the “quantum usability effect.” We turn now, therefore, to a series of observations derived from our data analysis that illustrate the major differences between traditional user testing and usability analysis conducted at the extremes of time and tests.

## **5. Observations on User Testing and the Quantum Usability Effect**

These observations underscore the primary differences between formal user testing, with its traditional goals of discovering as many usability flaws as possible, and user testing at “quantum extremes,” where the goal is simply to find something—anything—that will improve an interface as quickly as possible. They emphasize the point that such methods cannot be assessed using traditional methods of comparing usability evaluation methods, as it is unlikely that these methods will find the majority of all possible of usability flaws. Instead, these observations illustrate that if designers are to derive benefits from high-speed, low-cost user testing, usability evaluators perhaps should focus on finding something at the expense of finding everything.

*Observation 1) Within the Quantum Usability Effect, the number of tasks completed by user testers has no bearing on the evaluators’ ability to offer design recommendations.*

One of the more interesting findings of our data analysis was that our results showed no significant correlation between the number of tasks completed by our volunteer test users and the number of usability flaws or design recommendations we were able to make. This finding illustrates a significant difference between traditional user testing and testing at quantum extremes. For most usability evaluators, the number of tasks successfully completed by user testers is a valuable metric for assessing the usability of an interface (Nielsen, 1993). In our method of high-speed, low-cost user testing, the number of tasks successfully completed by our users was essentially irrelevant.

In our tests, it was not necessary for our users to complete tasks in order for us to learn something valuable. We were not terribly interested in how many tasks our users could or could not complete, nor did this affect the overall success of the method. What mattered for us and for our evaluations was that we saw enough of the interface, and found a suitable (non-zero) number of usability flaws which would allow us to make design recommendations. The tasks were what we asked our users to do in order to uncover interesting conclusions: they were the bait, and not the fish. Thus, when working within the quantum usability effect, whether or not user testers successfully complete their tasks has little or no bearing on the ability of usability evaluators to reach interesting conclusions and make design recommendations.

*Observation 2) Within the Quantum Usability Effect, it is neither necessary nor useful for evaluators to develop and administer tasks in a highly structured or systematic fashion.*

Throughout the course of our thirty-six trials of our method for high-speed user testing, we did not administer every task we developed in every user test (we administered, on average, three of every four developed tasks). While this can be partially explained by the fact that our thirty-minute evaluations included only ten minutes for user testing, running out of time was not the only reason for abandoning developed tasks. Given the pressures of testing at temporal extremes, we were often forced to play fast and loose with the “typical rules” of user testing, where evaluators proceed logically from task to task without intervening in the task completion process (Rubin, 1994; Maguire, 2001b; Paradowski and Fletcher, 2004). If one task was going particularly well (i.e. yielding many interesting findings), we sometimes focused on that task for several minutes, even if that meant we would not get to other tasks. On the other hand, if one task was not yielding particularly useful design recommendations, we had no qualms about jumping in and asking the users to move on to a different task, even before their original task was completed.

This approach to user testing is very different from traditional usability analysis methods, where tasks are carefully developed, highly structured, and systematically administered (Hackos and Redish, 1998). To ensure consistency from test to test, usability evaluators usually take pains to administer all the tasks they develop, in the same order, for each user. Since we were conducting only one user test, we were under no pressure to administer all of our tasks in a highly structured or systematic fashion. We were not trying to compare results across multiple tests, nor were we developing standardized tasks which could be posed to multiple users. Instead, our goal was to develop tasks which would produce non-zero findings in a very small amount of time.

It therefore was neither necessary nor advisable for us to administer all of our developed tasks systematically. If we were to be successful at finding usability flaws under these extreme conditions, we had to pick and choose our tasks during the user test (changing the order of developed scenarios, eliminating scenarios on the fly, modifying scenarios in progress, etc.) as we learned more about the interface we were evaluating. While such practices may be anathema in traditional user testing, they are the *sine qua non* of user testing at quantum extremes. In a ten-minute user test of an unfamiliar interface, rigidly administering every developed task could have been potentially disastrous. Our flexibility in developing and administering tasks helped us recover from unexpected errors or events, provided valuable alternatives that kept the evaluation process from bogging down, and helped us quickly develop many recommendations for design.

*Observation 3) Within the Quantum Usability Effect, administering a large number of quickly-developed tasks is more valuable for evaluators than administering a small number of carefully-developed tasks.*

One striking result from our data analysis was the finding that the more tasks we administered during our evaluations (whether or not they were completed by the users), the more usability findings we uncovered and the more design recommendations we made. This made it very important that we develop as many tasks as possible at the beginning of the evaluation, since the more tasks we developed, the more potential tasks we had to administer, and therefore the more results we were likely to find. With only ten minutes to analyze the site and develop representative tasks, however, we could not risk spending too much time developing tasks that

would prove unsuitable during the actual user test, especially since we had no prior knowledge of the site being evaluated.

We learned, therefore, that it was a better use of our time to develop as many tasks as possible, without paying too much attention to the question of whether they were suitable tasks for the evaluation. This is not to say that no thought was given to developing these tasks at all. As evaluators, we suggested tasks that made sense to us based on our understanding of the type of website being tested, which was possible even if we had never seen the specific website before. We did not have time, however, to discover in advance what might happen when our user testers attempted to complete each task. Unlike lawyers, who do not ask questions unless they already know the answer, we consistently posed tasks to our users without knowing where in the website this task would take us.

This is very different from traditional user testing, where usability evaluators carefully pick representative tasks that highlight specific areas of interface in order to evaluate specific usability concerns (Nielsen, 1994; Rosson and Carroll, 2001; Maguire, 2001a). In our evaluations, we could not afford to spend time carefully selecting tasks with a particular goal in mind, because we did not have enough data about the interface we were evaluating. It was impossible for us to tell in advance which tasks would pay off in an intriguing way, thereby giving us rich insights, and which tasks would be a waste of our extremely-limited time. Given the constraints we faced, it was safer to develop a number of tasks that covered a wide variety of topics than it was to risk spending even two or three minutes developing any one task. In addition, having many tasks from which to choose helped us quickly administer many tasks, thereby increasing our chances of finding problems during the user test.

When conducting user tests at quantum extremes, therefore, we consistently found that having a bunch of quickly-developed tasks enabled us to catch a variety of usability problems, ranging from simple interface flaws to complex reasoning issues. Although these tasks may not have been perfect, they frequently yielded valuable findings, similar to the techniques of rapid paper prototyping (Hall, 2001; Rudd et al, 1996; Rettig, 1994; Virzi, 1989). When planning and executing user tests under such conditions, therefore, we found that a large number of quickly-designed tasks was more likely to provide a greater return on investment than a smaller number of carefully-designed tasks.

*Observation 4) Within the Quantum Usability Effect, identifying a “sweet spot” of errors that will guarantee a non-zero number of design recommendations is more important for evaluators than attempting to find the majority of usability flaws.*

As discussed above, our goal in these evaluations was to find a non-zero number of usability flaws in a very small amount of time. We could not waste time on tasks that would not yield any useful design recommendations, nor could we tell in advance which tasks would cause problems and which would not. The trick to coping with these challenges was to spend the ten-minute user test casting around, constantly trying different tasks until we found tasks that produced valuable findings. What mattered was whether we found things to say; the more tasks we administered, the more of the interface we saw, and the more likely it was that we would be able to make useful design recommendations.

The reason this approach succeeded was that we were specifically not trying to maximize the number of usability flaws and design recommendations. Unlike traditional usability evaluators, we were purposely looking for a “sweet spot” of usability errors, where we would be virtually guaranteed a non-zero number of findings. In searching for this “sweet spot,” we purposely skipped tasks that looked like they would not yield many findings, even if they might have produced something eventually had we had more time. Given the time pressures we were under, it was far more important for us to test several tasks that did not take too long to complete, than fewer tasks that took longer yet might not result in many findings.

This process of searching for a “sweet spot” of usability errors makes sense when viewed as part of a “berry-picking” or “low-hanging” fruit process. If one cannot tell in advance where the pickings will be most productive, it makes the most sense to keep trying several different sections until one finds the richest spot. Once we found that “sweet spot,” we knew that by focusing on just that one area, we could essentially dismantle the interface, peeling back the veneer of even the most polished application to reveal numerous usability flaws underneath.

Even though we expected this to happen, we were still surprised when it happened thirty-six out of thirty-six times. We were constantly amazed by how rich this approach to user testing turned out to be. If we had tried to maximize our numbers by systematically searching for every usability flaw, we would likely have run out of time before finding a similar “sweet spot” of usability errors and never have found such rich results. When working at the quantum extremes of user testing, therefore, trying to find results that are non-zero is much more likely to produce valuable results than systematically trying to find all usability errors with an interface.

*Observation 5) Within the Quantum Usability Effect, the results of the usability evaluation should be validated internally while the method is in progress, and not externally, after the evaluation is complete.*

Our overall goal in these evaluations was to provide valuable feedback to site designers that would allow them to improve the usability of their websites. As noted above we could not use traditional external methods to validate the usability of a site by giving it any kind of score or passing grade. We could not say that these results held true over multiple tests with many different users. We could not say that our results represented the majority or even the most significant usability flaws found with the sites tested. We certainly could not say that in only thirty minutes we had found all the problems the sites’ designers needed to worry about! These constraints made our approach very different from traditional user testing, where results are formally and externally validated after the evaluation is complete, either by comparing the results to those from other evaluations, or by conducting more tests (Gray and Salzman, 1998). Since we could not do that, all we could do is gather a number of usability concerns for our participants to consider, along with some suggestions for how they might be fixed. Ultimately, we were simply looking for something our participants could use to increase the overall usability of their sites.

Our approach to high-speed user testing, therefore, should be seen as a kind of formative evaluation, drawing heavily on informal, qualitative methods (Twidale, 1993). Its role is to suggest aspects of the interface or the user experience that may be likely to cause users

difficulties, and to inspire ideas for design solutions by elaborating the underlying causes of these difficulties or confusions. It is important to clarify that this method cannot serve as a summative evaluation method, to validate or even to effectively compare one interface with another. Given its formative nature it necessarily has to concentrate on finding usability problems: things that seem to be wrong with the interface, or that could be improved. This focus on flaws means that we, as evaluators, often had a rather inverted view of usability problems. For us, the absence of usability flaws (or our failure to detect any flaws in the available time) was a failure, whereas their detection, elaboration, and possible correction was a success, particularly in the public, pedagogical context of our studies. While this approach runs the risk of making the results appear rather negative, our goal was not to detract from the overall quality of the interfaces we studied (many of which were award-winning websites), but rather to provide helpful, constructive criticism to improve their overall usability.

In many ways, this is similar to a situation where athletes ask their coach to tell them what to do to improve their game. A coach that says, “Nothing is wrong, you are doing just fine!” is not doing his or her job correctly. In our evaluations, it was equally important for us to arrive at a non-zero number of usability findings and design recommendations. At the same time, our participants had to believe that what we told them to work on was actually a problem, and that by implementing our recommendations they would make their sites better and not worse. Given the limitations of our method, therefore, it was our participants who validated the usability findings and design recommendations. If they did not believe in our results, the evaluation had been a waste of time, ours and theirs. If they did believe in our results, then they were more likely to take the required steps to improve the usability of their system, starting—but not stopping—with the recommendations we made.

Even though our findings were necessarily tentative, they could as we noted above, be externally validated by appeals to HCI theory, published studies reporting related effects, as well as by the comments and reactions of the audience and test participants. In some cases, we were confident in claiming that we uncovered a real flaw that needed fixing, just as if we had conducted a study using inspection methods such as heuristic evaluation or a cognitive walkthrough (Nielsen and Mack, 1994). Indeed, there are many parallels between our high-speed usability tests and inspection methods, which we intend to elaborate on in future work. In other cases, our user study may have uncovered an interesting confusion, but we lacked the external validation necessary to assert that the problem deserved an immediate design fix. In such cases, the study acted like a rapid pilot test for a larger, more systematic experiment, uncovering interesting issues for the larger study to investigate in more depth.

Over the years while we have been developing and testing this method, we have observed designers take the results of our extremely-short user test and use those results to guide complete redesigns of their sites—often beginning those redesigns while our analysis was in progress. We have had designers contact us years later and comment on how participating in our thirty-minute evaluation changed their approach toward user testing and usability analysis. Once we, as evaluators, accepted that the constraints of testing at quantum extremes meant we could not validate our results after the fact using traditional methods, we quickly realized the long-term benefits of asking our participants to validate our results in progress.

## 6. Conclusions and Implications for the Future of User Testing

This article discussed the difficulties of performing usability analyses by conducting only one user test in a very short amount of time. It presented several observations about how usability evaluators need to bring different approaches to usability analysis when working at these “quantum extremes” of user testing. It demonstrated not just the dangers of testing at extremes, but also the benefits that can accrue to the evaluators and designers.

When pursuing extremely discounted usability engineering, the relevant questions tend mostly to revolve around the issue of “satisficing:” How much effort do usability evaluators need to expend before they can do some positive amount of good? From a scientific perspective, the answer to this question involves numerous tests, extensive data gathering, and rigorous data analysis—all necessary elements that will allow researchers to draw generalizable conclusions that will be valid in other situations. From an engineering perspective, however, the answer to the “satisficing question” rarely involves gathering sufficient data to address universal usability truths; most designers are simply concerned with making a positive impact on the usability of their systems (Wixon, 2003).

What we learned from our study was that a focus on the practical, but not necessarily scientific, side of usability analysis can hold true even at the quantum extremes of user testing, where evaluators conduct only one user test in a very short amount of time. Our method for high-speed, low-cost user testing provided us with a wide-mesh fishing net that allowed us to catch big fish quickly and cheaply. Hearing of our method for the first time, a reader might suspect that our net was so wide-mesh and that we fished for such a short period of time that we would catch nothing other than the odd, non-fishy tree branch. By the end of our tests, however, our unusual method of fishing had caught more than 500 fish in only 18 hours of work. Of course we missed hundreds of other fish, and many of them were little fish that our net was too wide-mesh to catch. But we were not interested in catching every fish in the sea, just as many fish as could, provided we caught them quickly and easily. In the end, in order to prove that our net could be used for fishing, all we needed to catch was one fish, and we got one, and more, every time.

Unlike traditional user testing where evaluators are trying to be exhaustive and find all problems with an interface, we were not concerned with finding all problems, most problems or even most major problems. We were simply not concerned with being thorough; rather we were interested in seeing how quickly we could find something—anything—that would be useful for designers. Our experiment demonstrated that even when discount usability engineering is taken to extremes, it still produces results that are better than nothing.. When faced with seemingly overwhelming obstacles to formal user testing (not enough time, only one test possible), researchers who understand these limitations and work with them, instead of against them, will succeed in producing useful results. Against all odds, they will discover that even when fishing with a very wide net for a very short period of time, they will not go away hungry. This has several important consequences with implications for the future of usability analysis.

First, our findings open up a relatively unexplored area of the user testing continuum. Many usability evaluators have noted the distinctions between doing academic research and pragmatic usability, as well as the huge gap between the amount of effort required for traditional, formal

user testing and quick, informal design studies (Sauro, 2004). While we are not saying that our method is an improvement on formal, scientific, summative evaluations, we are saying we speculated on and investigated an extreme along the user testing continuum. Our goal in doing so was not necessarily to advocate for that extreme position, but to emphasize the existence of this continuum and to note the differences of user testing at extremes. As such, our approach can be used not to advocate one method over other methods, but to explore the total design space of usability testing methods. Methods such as the one we used in this study, therefore, become analysis tools or debating points that could be used as a basis for designing other minimalist methods, providing a set of examples that demonstrate a different way of looking at the issue from much of the current debate about the validity of usability evaluation methods.

Second, our findings illustrate that high-speed, low-cost discount usability engineering can produce useful results, even at extremes of time and tests. A method such as the one we used here allows for the quick identification and fixing of certain “obvious” problems, where all participants agreed that a problem was one that others would likely encounter and needed to be fixed immediately. In these cases, it seemed to be acceptable, and indeed expedient, to all involved that it would be in everyone’s best interest for the designers to quickly fix that problem and move on. Moreover, even in other less clear cut situations, unexpected results observed in such a rapid study can serve as an incentive for more systematic future investigation. This immediacy of results is a powerful way of promoting a culture of “test early and test often,” and advocating the importance of usability evaluation pervading the design process. This approach also, of course, can be dangerous: while a sample of one can highlight potential problems, it is not a substitute for multiple tests, and more testing is usually necessary in order to determine whether problems are indeed widespread enough to warrant design changes. Evaluators using extreme discount usability engineering methods, therefore, should be very explicit about what they can and cannot guarantee.

Third, our findings raise serious questions about the circumstances in which such methods should or should not be employed. Evaluators must be able to justify the level of rigor required for each evaluation they undertake. They must be able to recognize the circumstances where their results need not be scientifically valid, but merely good enough for design purposes. To do so, they need a firm understanding of the five observations discussed above that explain the challenges present in conducting user tests at quantum extremes. It would not be advisable for unskilled evaluators to attempt to substitute informal, extremely-discounted usability methods for more scientific, formal evaluation methods. In our experiments, we often had to explain why we felt our results were valid in each circumstance, and in doing so we drew upon many years of experience with user testing to draw believable claims from such small amounts of data. Arriving at these sorts of conclusions is not easy, and not just anyone will be able to run off and conduct tests using this method; there could be serious consequences that backfire on the unskilled evaluator. It would be worth exploring in future studies the circumstances in which such low-cost, high-speed methods might do more harm than good.

Fourth, and finally, we believe that our work in this area has the potential to be integrated into a number of other evaluation and development processes, such as:

- *Extreme Prototyping*. If our method for extreme user testing could be integrated with Extreme Prototyping (Beck, 2000), it could help enable more rapid design iterations. Most of the evaluation work in extreme prototyping has so far concentrated on testing evolving code to see if it matches specifications for bug free performance. We believe that with our evaluation methodology, similar tests for usability performance would be both valuable and feasible.
- *Open Source Usability*. There is a growing discussion about the challenges of incorporating better usability analysis into open source software development (Nichols and Twidale, 2003). One challenge lies in enabling volunteers to participate in the evaluation process in a manner analogous to the way code development is federated and integrated. Extreme usability tests offer interesting possibilities, enabling low-cost participation by several volunteers running a handful of short usability tests and combining the results: an approach that would fit very nicely with the “nightly-build” approach to iterative open source software development.
- *Education and Evangelization*. One reason we developed our method was that we believe it has a role to play in the instruction of future computer scientists and other HCI practitioners who hopefully will become more attuned to working with usability experts in iterative project design and development. We also believe it has a role outside of the classroom in convincing project stakeholders that usability is important, cost effective, and can be integrated into time-critical development paths. High-speed, low-cost methods user testing can be very helpful in spreading the word about the importance of usability.
- *Rapid Evaluation Prototyping*. We believe that our method of user testing can be valuable even in cases where a thorough, systematic, and scientific evaluation is planned with substantial numbers of subjects and sessions of significant. The availability of low-cost, high-speed pilot tests can help in the planning and refinement of the more formal evaluation method. While most evaluators run a pilot study, we believe it would be interesting to explore the potential of running a sequence of very rapid pilot studies, with small iterative refinements to the method, in preparation for more formal user testing.
- *Usability Inspection Methods*. Lightweight evaluation methods, such as with heuristic evaluation and cognitive walkthroughs, that allow testing without users have been shown to catch certain categories of errors rapidly and at low cost (Nielsen and Mack, 1994). It would be interesting to explore whether the combination of these methods with extreme user testing would produce even richer data, in exchange for relatively little extra effort, time, or expense.
- *Extreme Longitudinal Evaluations*. If one assumed a rapid prototyping cycle where high-speed, low-cost user tests were conducted once every couple of days, one would create a dataset of the results of a sequence of tests over time. Since each study would have been done on different versions of the software and may have involved different tasks or emphases, the results would not be directly comparable, as in traditional scientific studies. Nevertheless, it would be interesting to study whether some valuable trends or issues would be derived from these data. Our own experiences of analyzing the finding from thirty-six completely different applications lead us to believe that a similar analysis of a single evolving application is likely to be fruitful, provided that appropriate analytic methods are developed.

In conclusion, lest anyone misunderstand, we are not saying that any user test, no matter how poorly done, is better than no user test at all; if done badly, the results of one test can be just plain wrong and ultimately damaging to the design process. We are saying, however, that when one fast user test is conducted by evaluators with a full understanding of the quantum usability effect, the results can be absolutely astounding. It is simply amazing how quickly evaluators learn something when they are not trying to learn everything. We firmly believe that, when properly leveraged, the results of these and similar high-speed, low-cost user testing methods could have radical implications for the future of usability analysis. Usability evaluators who understand how to cope with the unique conditions of the quantum usability effect will be better prepared to employ discount usability engineering techniques at the extremes of time and tests.

## References

- Bauersfield, K., Halgren, S., 1996. "You've got three days!" Case studies in the field techniques for the time-challenged, in: Wixon, D., Ramey, J. (Eds.), *Field Methods Casebook for Software Design*. Wiley, New York, pp. 177-196.
- Beck, K., 2000. *Extreme Programming Explained: Embrace Change*. Addison-Wesley, New York.
- Bias, R. G., Mayhew, D. J. (Eds.), 1994. *Cost Justifying Usability*. Academic Press, Boston.
- Desurvire, H. W., 1994. Faster, Cheaper!! Are usability inspection methods as effective as empirical testing?, in: Nielsen, J. Mack, R. (Eds.), *Usability inspection methods*. Wiley, New York, pp. 173-202.
- Dicks, R.S., 2002. Mis-usability: On the uses and misuses of usability testing, in: *Proceedings of SIGDOC'02*. ACM Press, New York, pp. 26-30.
- Donahue, G., 2001. Usability and the bottom line, *IEEE Software*, January/February (2001), 31-37.
- Dumas, J., 2002. User-based evaluations, in Jackie, J., Sears, A. (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, Lawrence Erlbaum Associates, Inc., Mahwah, NJ, pp.1093-1117.
- Faulkner, L., 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments and Computers* 35(3), 379-383.
- Gray, W. D., Saltzman, M. C., 1998. Damaged Merchandise? A review of experiments that compare usability evaluation methods. *Human Computer Interaction* 13, 203-261.
- Hackos, J.T., Redish, J.C., 1998. *User and task analysis for interface design*. John Wiley & Sons, Inc., New York.
- Hall, R., 2001. Prototyping for usability of new technology. *Int. J. Human-Computer Studies* 55, 485-501.
- Jeffries, R., Miller, J.R., Wharton, C., Uyeda, K.M., 1991. User interface evaluation in the real world: A comparison of four techniques. in: *Proceedings of CHI'91 Human Factors in Computing Systems*, ACM Press, New York, pp. 119-124.
- Jordan, P., Thomas, B., Weerdmeester, B., McClelland, I. (Eds.), 1996. *Usability Evaluation in Industry*. Taylor & Francis, Ltd., London.

Karat, C.M., Campbell, R.L., Fiegel, T., 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. Proceedings of CHI'92 Human Factors in Computing Systems, ACM Press, New York, pp. 397-404.

Landauer, T.K., Nielsen, J., 1993. A mathematical model of the finding of usability problems. Proceedings of CHI'93 Human Factors in Computing Systems, ACM Press, New York, pp. 206-213.

Maguire, M., 2001a. Context of use within usability activities. *Int. J. Human-Computer Studies* 55, 453-483.

Maguire, M., 2001b. Methods to support human-centered design. *Int. J. Human-Computer Studies* 55, 587-634.

Marty, P.F., Twidale, M.B., manuscript under review. Usability@90mph: Maximizing the benefits of user testing demonstrations.

Marty, P.F., Twidale, M.B., 2004. Lost in gallery space: A conceptual framework for analyzing the usability flaws of museum websites. *First Monday*, 9 (9), available online at [http://www.firstmonday.org/issues/issue9\\_9/marty/index.html](http://www.firstmonday.org/issues/issue9_9/marty/index.html), accessed September 15, 2004.

Nichols, D.M., Twidale, M.B., 2003. The usability of open source software. *First Monday*, 8 (1), available online at [http://firstmonday.org/issues/issue8\\_1/nichols/index.html](http://firstmonday.org/issues/issue8_1/nichols/index.html), accessed September 15, 2004.

Nielsen, J., 1993. *Usability Engineering*. Academic Press, Boston.

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

Nielsen, J., 1994b. Estimating the number of subjects needed for a thinking aloud test. *Int. J. Human-Computer Studies* 41, 385-397.

Nielsen, J., Mack, R.L. (Eds.), 1994. *Usability Inspection Methods*. John Wiley & Sons, New York.

Nielsen, J., Philips, V.L., 1993. Estimating the relative usability of two interfaces: heuristic, formal and empirical methods compared. Proceedings of CHI'93 Human Factors in Computing Systems, ACM Press, New York, pp. 214-221.

Paradowski, M., Fletcher, A., 2004. Using task analysis to improve usability of fatigue modeling software. *Int. J. Human-Computer Studies* 60, 101-115.

Rettig, M., 1994. Prototyping for tiny fingers. *Communications of the ACM* 37 (4), 21-27.

- Rosson, M.B., Carroll, J. M., 2001. Usability Engineering: Scenario-Based Development of Human-Computer Interaction. Morgan Kaufmann Publishing, San Francisco.
- Rubin, J., 1994. Handbook of User Testing. John Wiley & Sons, New York.
- Rudd, J., Stern, K., Isensee, S., 1996. Low vs. high-fidelity prototyping debate. Interactions 3 (1), 76-85.
- Sauro, J., 2004. Premium usability: getting the discount without paying the price. Interactions 11 (4), 30-37.
- Siegel, D., 2003. The business case for user-centered design: increasing your power of persuasion. Interactions 10 (3), 30-36.
- Spool, J., Schroeder, W., 2001. Testing web sites: Five users is nowhere near enough. In CHI 2001 Extended Abstracts, ACM Press, New York, pp. 285-286.
- Thomas, B., 1996. Quick and dirty usability tests, in: P. Jordan, B. Thomas, B. Weerdmeester, I. McClelland (Eds.), Usability Evaluation in Industry. Taylor & Francis, Ltd., London, pp. 107-114.
- Twidale, M. B., 1993. Redressing the balance: the advantages of informal evaluation techniques for Intelligent Learning Environments. Journal of Artificial Intelligence in Education 4 (2/3), 155-178.
- Virzi, R.A., 1989. What can you learn from a low-fidelity prototype? Proceedings of Human Factors Society (Santa Monica, CA), pp. 224-228.
- Virzi, R.A., 1992. Refining the test phase of usability evaluation: how many subjects is enough? Human Factors 34, 457-468.
- Waes, L., 2000. Thinking aloud as a method for testing the usability of websites: the influence of task variation on the evaluation of hypertext. IEEE Transaction on Professional Communication 43, 279-291.
- Wildman, D., 1995. Getting the most from paired user testing. Interactions 2 (3), 21-27.
- Wixon, D., 2003. Evaluating usability methods: why the current literature fails the practitioner. Interactions 10 (4), 28-34.
- Wixon, D., Ramey, J. (eds.). Field Methods Casebook for Software Design. John Wiley & Sons, New York, 1996.
- Woolrych, A., Cockton, G., 2001. Why and when five test users aren't enough, in: Vanderdonckt, J., Blandford, A., Derycke, A. (Eds.), Proceedings of IHM-HCI 2001 Conference Vol. 2. Cépadèus, Toulouse, France, pp. 105-108.

## Tables

|                        |                           |                        |                                 |
|------------------------|---------------------------|------------------------|---------------------------------|
| <i>Tasks Developed</i> | <i>Tasks Administered</i> | <i>Tasks Completed</i> | <i>Findings/Recommendations</i> |
| 151                    | 119                       | 83                     | 500+                            |

Table 1: Total Results after 36 Trials of Method

|                        |                           |                        |                                 |
|------------------------|---------------------------|------------------------|---------------------------------|
| <i>Tasks Developed</i> | <i>Tasks Administered</i> | <i>Tasks Completed</i> | <i>Findings/Recommendations</i> |
| 4.2                    | 3.3                       | 2.3                    | 10-15                           |

Table 2: Average Results of one Thirty-Minute Evaluation

| Test | Number of Volunteers | Tasks Developed | Tasks Administered | Tasks Completed | Findings / Recommendations |
|------|----------------------|-----------------|--------------------|-----------------|----------------------------|
| 01   | 1                    | 3               | 3                  | 3               | 5-10                       |
| 02   | 2                    | 5               | 3                  | 3               | 10-15                      |
| 03   | 1                    | 4               | 3                  | 2               | 20-25                      |
| 04   | 2                    | 3               | 3                  | 3               | 10-15                      |
| 05   | 1                    | 5               | 3                  | 1               | 10-15                      |
| 06   | 2                    | 3               | 3                  | 2               | 5-10                       |
| 07   | 1                    | 4               | 3                  | 2               | 15-20                      |
| 08   | 2                    | 6               | 5                  | 5               | 10-15                      |
| 09   | 1                    | 4               | 4                  | 1               | 15-20                      |
| 10   | 1                    | 8               | 5                  | 3               | 15-20                      |
| 11   | 2                    | 5               | 4                  | 4               | 10-15                      |
| 12   | 1                    | 3               | 3                  | 2               | 10-15                      |
| 13   | 2                    | 4               | 3                  | 3               | 10-15                      |
| 14   | 1                    | 3               | 2                  | 1               | 10-15                      |
| 15   | 2                    | 6               | 3                  | 2               | 5-10                       |
| 16   | 1                    | 4               | 3                  | 1               | 10-15                      |
| 17   | 1                    | 6               | 4                  | 4               | 5-10                       |
| 18   | 1                    | 4               | 3                  | 1               | 15-20                      |
| 19   | 2                    | 5               | 4                  | 4               | 15-20                      |
| 20   | 1                    | 1               | 1                  | 1               | 10-15                      |
| 21   | 2                    | 4               | 4                  | 3               | 10-15                      |
| 22   | 1                    | 5               | 2                  | 1               | 10-15                      |
| 23   | 2                    | 4               | 2                  | 2               | 5-10                       |
| 24   | 1                    | 3               | 3                  | 2               | 10-15                      |
| 25   | 2                    | 4               | 4                  | 2               | 10-15                      |
| 26   | 2                    | 4               | 3                  | 0               | 5-10                       |
| 27   | 1                    | 5               | 5                  | 3               | 20-25                      |
| 28   | 2                    | 4               | 4                  | 3               | 15-20                      |

|    |   |   |   |   |       |
|----|---|---|---|---|-------|
| 29 | 1 | 4 | 4 | 2 | 15-20 |
| 30 | 2 | 6 | 5 | 4 | 15-20 |
| 31 | 1 | 1 | 1 | 1 | 5-10  |
| 32 | 1 | 5 | 5 | 3 | 10-15 |
| 33 | 2 | 5 | 2 | 2 | 10-15 |
| 34 | 1 | 3 | 3 | 2 | 15-20 |
| 35 | 2 | 3 | 2 | 1 | 10-15 |
| 36 | 1 | 5 | 5 | 4 | 15-20 |

Table 3: Individual Results for each of 36 Trials of Method