# A Comparison of Methods for Eliciting Post-Task Subjective Ratings in Usability Testing

Donna P. Tedesco
Fidelity Investments
donna.tedesco@fmr.com

Thomas S. Tullis
Fidelity Investments
tom.tullis@fmr.com

## ABSTRACT

Five methods for eliciting subjective ratings after each task in a usability test were evaluated. The methods included simple Likert scales as well as a technique derived from Usability Magnitude Estimation. They were tested in a large-scale online study in which participants performed six tasks on an Intranet site. Performance data for the tasks reflected the same pattern as all of the subjective ratings. All five methods yielded significant differences in the subjective ratings for the tasks. A sub-sampling analysis showed that one method yielded the most reliable results at the small sample sizes typical of usability tests.

## Introduction

Usability professionals sometimes elicit post-task subjective ratings from usability test participants to understand the ease or difficulty of use of a product for each task or scenario. Some professionals use them to collect specific quantifiable data regarding individual tasks. Others argue that the data seems useless because of the low number of participants, and they use post-task ratings more as a springboard to get participants speaking about their feelings and experiences during the task. Whether using post-task ratings for qualitative data, quantitative data, or both, the goal is universally clear; to further understand the usability issues of the product surrounding a task, or in how a set of tasks relate to each other. But there is little consensus or consistency in whether or not usability professionals use post-task ratings to begin with, let alone how the ratings are used. This is likely because not much research has been conducted on the value of post-task ratings.

Lewis (1991) uses three Likert scale questions called the After-Scenario Questionnaire (ASQ) which ask for the participants' level of satisfaction with the ease of completing the task, the amount of time it took to complete the task, and the quality of support information (e.g. online help) to complete the task. He found that the ASQ was correlated with task completion data and that the ASQ suggests a useful measure of usability. Albert & Dixon (2003) present the use of expectation ratings for usability testing, in which users are asked beforehand how easy or difficult they expect a task to be and are asked again after the task how easy or difficult it actually was. They suggest that understanding how a task actually was relative to the participant's expectations for it allows usability professionals to clearly identify necessary areas of improvement for the product. McGee (2004) uses methods called Usability Magnitude Estimation and Master Usability Scaling to allow participants to rate the usability of a product after a task using a ratio rating relating it to previous tasks. He found that this is a more accurate and revealing measure than typical Likert scales, resulting in usability scores or percentages for particular tasks as well as products overall. This method is used successfully by some companies today.

The study described in this paper was conducted to compare five post-task rating methods, including those mentioned above. The goals of this study are very similar to those of a study conducted by Tullis & Stetson (2004), in which several post-*study* rating methods were compared. Our study aimed at understanding whether any one method for administering post-task subjective ratings might better indicate usability issues than the others.

## Method

The study was run as an online, self-administered usability study. We chose this because we could easily obtain large numbers of participants in very little time and for little incentive (users would be entered into a drawing for a gift check). The study was completed by 1,131 employees of Fidelity Investments. Users were notified of the study in a company-wide morning message posted for two days in January 2005 and participated on a voluntary basis by clicking on a link in the posting. The link brought users to a 'welcome' page, which gave them general instructions and "background" on the study, as they thought that the purpose of the study was to evaluate the usability of a website. Upon continuing, participants were randomly brought to 1 of 5 conditions of post-task ratings (described later in this section). For all conditions, users were given the same set of 7 tasks to complete on an internal website which allows employees to lookup contact and organizational information for other employees. Employees have varying levels of experience with this site, but from the authors' knowledge of the company as well as open-ended comments from users of the test, it is evident that most users are very familiar with the site and use it frequently. Tasks were purposely chosen to vary in difficulty to offset any floor effects for newer users or, more importantly, ceiling effects for experienced users. Participants were presented with tasks one at a time. This was done by having two windows launched on the screen; the top window presented the task and ratings, and the bottom window presented the website, as shown in Figure 1.

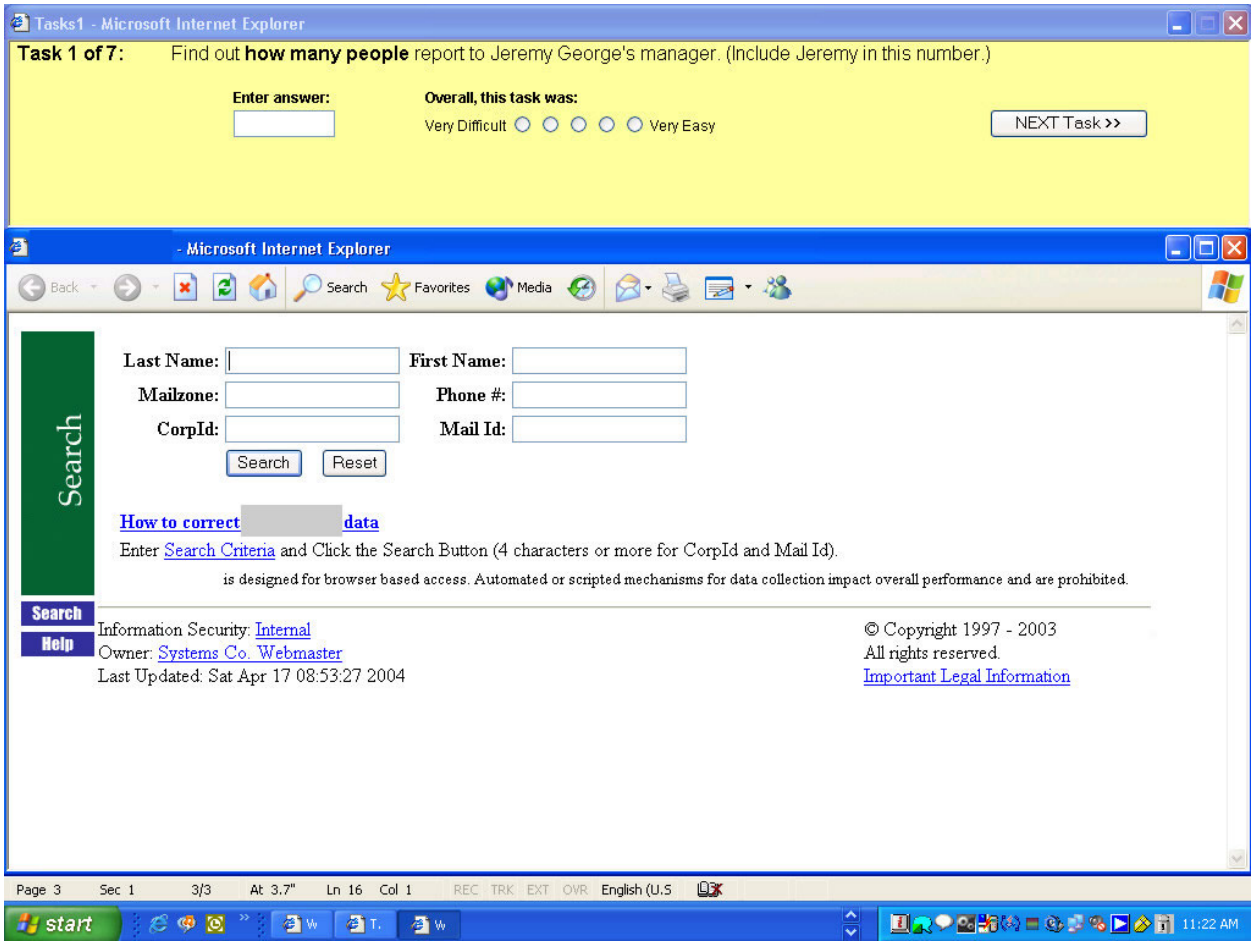


**Figure 1: When starting the study, a top (tasks & ratings) and bottom (site) window is launched on the user's desktop.**

For each task, the user was instructed to find the answer using the website displayed in the bottom window, type their answer to the task, rate the task, and click a button to go to the next task. Answers, ratings, and time to complete the task were captured to a database. At the end of the test, users were asked to complete a post-study questionnaire which consisted of the standard SUS (System Usability Scale) survey as well as an open-ended comments/suggestions field.

The seven tasks presented to the participants are listed below. Although somewhat cryptic to the outside reader, the jargon used in these tasks (acronyms, terms, etc.) was common company knowledge.

1. How many people report to Bill George's manager? (Include Bill in this number.)
2. You need to return a notebook to someone but only know that he works on the 4th floor of 245 Summer St (building code V, floor 4), and that his manager's name is Tom. Find this man and enter his corporate ID.
3. You remember talking with someone named John and want to contact him, but don't remember his last name. You only know that his last name starts with S. You also remember that he works at 500 Salem St. in Smithfield (Building code SM) and works in FIIS. Find this man and enter his corporate ID.
4. You know a woman who works in FISC whose first name is Edyta and you need to call her. What are the last four digits of her phone number?
5. You want to send a meeting agenda to a woman who sits next to your friend Martha Jones. You call Martha to find out who it is, but she's not there. Desperate to find this woman's name, you turn to this website. Find the person and enter her corporate ID.
6. You get to your desk and see that someone just called you, but didn't leave a message. You're curious to know who called and look at the call log. Your phone shows that the phone number was 8175559717. You recognize this as possibly a Fidelity internal number. Find out who this was and enter his/her corporate ID.
7. You have worked with a man named Jeremy Bennett and his manager Paul. What are the last four digit's of Jeremy's phone number?

***Conditions***

Upon beginning the study, all participants were randomly directed to 1 of 5 conditions, making this a between-subjects study design. Each condition was a different method for eliciting post-task subjective ratings and is described below.

- **Condition 1**

  "Overall, this task was:

  Very Difficult ○ ○ ○ ○ ○ Very Easy" (Circles are representative of 5 radio buttons)

  This is our usability group's standard post-task question. It is based on what seems to be most frequently used in the usability field.

- **Condition 2**

  "Please rate the usability of the site *for this task:*

  Very Difficult to Use ○ ○ ○ ○ ○ Very Easy to Use"

  This question is potentially another way to ask the question given for Condition 1. Specifically, we wondered if there would be a difference in the way users interpreted and answered the questions in these two conditions. Would rating how easy or difficult *the site was* for a task be different from rating how easy or difficult *the task* was? There may be a subtle difference in our minds, but we wanted to bring that question to the users.

- **Condition 3**

  "Overall, I am satisfied with the ease of completing this task:

Strongly Disagree ▯ ▯ ▯ ▯ ▯ Strongly Agree

Overall, I'm satisfied with the amount of time it took to complete this task:

Strongly Disagree ▯ ▯ ▯ ▯ ▯ Strongly Agree"

These are 2 of the 3 questions used in Lewis' ASQ (1991). The third question of the ASQ asks about the support information such as online help. We found this question to be irrelevant given the study goals and type of interface used and therefore did not use it.

- **Condition 4**

  (Before doing all tasks): "How difficult or easy do you expect this task to be?

  Very Difficult ▯ ▯ ▯ ▯ ▯ Very Easy"

  (After doing each task): "How difficult or easy did you find this task to be?

  Very Difficult ▯ ▯ ▯ ▯ ▯ Very Easy"

  This is a 'before and after' or expectation rating system presented by Albert and Dixon (2003). The premise of the system is that rather than using an absolute rating, more important information can be extracted from looking at the relationship between a user's expectations for a task, and how the task actually was in relation to that initial expectation. For this condition, users were first brought to a page asking them to rate their expectations for all of the tasks they will be performing on the site. They were told that this rating should truly be how they *expect* the task to be, rather than *want* it to be, and that the expectation should reflect their assessment of the site's expected usability rather than their own skills. They were also asked not to "cheat" by performing the tasks ahead of time. Users then performed the task in the same manner as the other conditions and entered their "after" rating once they've completed each task.

  Note that the "after" rating wording is yet another variation of what was asked in Conditions 1 and 2.

- **Condition 5**

  "Please assign a number between 1 and 100 to represent how well the Website *supported you* for this task.

  Remember—**1** would mean that the site was *not at all* supportive, and completely unusable. 100 would mean that the site was *perfect*, and would require absolutely no improvement.

  Enter Rating:"

  This condition was significantly modified through iterations in the study planning. We began by wanting to evaluate the Usability Magnitude Estimation technique (McGee, 2004), with its roots in a method originally used in psychophysics that has been adapted for use in usability testing. The theory behind Magnitude Estimation is that ratings are much more accurate if they're not constrained by boundaries but rather are values given as ratios; i.e., rather than having a user choose a number between 1 and 5, you would allow the user to come up with an arbitrary number to assign to the usability of the product for a task, and for the next task have them assign another number as a factor or multiple of the previous number. So, if task 1 is rated as a 100 for usability and task 2 seems three times as easy with the product, the user would rate the second task as a 300. Each user's ratings can then be normalized for that person and across participants, and the data can then be transformed so that it results in an overall usability score for each task. For more information on Usability Magnitude Estimation, see McGee's paper (2004).

  Magnitude estimation is meant to be used in person, so that it could easily be explained and demonstrated to a usability participant. We found that adapting this method to an online study was very difficult. We tried breaking up the instructions and information into a wizard of multiple steps, and implemented practice exercises in our first versions of the online study, but found that the concept was too cumbersome, long, and confusing for people to understand and attempt. As a result, we decided to modify this condition to something similar, but more easily understood.

For the modified version we still used the idea of proportional rating, but gave users anchors of 1 and 100. Participants were given an instructional page explaining how they should rate each task. The instructions were as follows:

> "After each task, you will rate how well the website supports you in completing that task. You will assign this rating on a scale of 1 to 100. A rating of 1 means that the site was not at all supportive, i.e. completely unusable and should drastically change to complete that task easily. A rating of 100 means that the site was perfect, i.e. nothing should be improved on the site to make that task easier to accomplish. If you think the site falls somewhere between the two statements above, assign a number accordingly. Because the scale is large, you are free to rate the usability of tasks proportionally. For example:
>
> If the site hardly supports you (seems somewhat difficult to use) for Task 1, let's say you rate it a 20.
>
> For Task 2, suppose the site supports you much better — and if you were to estimate, perhaps it feels roughly about 4 times as easy to accomplish than the last task. Then you would give Task 2 an 80.
>
> Try to think proportionally in this way when rating tasks. There are no right or wrong ratings...just go with your instinct and estimate."

We found that pilot participants found this easier to understand and were more willing to attempt the online study.

## Analysis and Results

As mentioned before, 1,131 employees of our company participated in the study. The random assignment of participants to conditions resulted in 210, 230, 244, 227, and 221 participants for Conditions 1-5, respectively. Because we were testing a live site, the answer to Task 7 had changed and was no longer available on the second day of testing. We therefore received meaningless data for this task and it was discarded, leaving six tasks for analysis.

We first compared the results of the two Likert scales in Condition 3 (the ASQ) using a t-test and found that there were no significant differences in responses between the two questions; i.e. participants were likely to give the same or a similar rating to both questions. From that point forward we took the average of the two questions when comparing it to other conditions.

For the expectation ratings in Condition 4, there was no apparently simple way to compare the before/after ratings with the other conditions. However for the purpose of comparing conditions in this study, we only used the "after" ratings (users' actual experiences with the site, rather than expected).

To transform the ratings for each condition in a way that could be compared across all conditions, we converted ratings to percentages. For the Likert scales in conditions 1-4, this meant converting the 5-point scale to their equivalent percentages, i.e., 20%, 40%, 60%, 80%, and 100% respectively. When surveying the ratings data for Condition 5 (the 1-100 scale), we found that users' tended to treat this like any other scale regardless of the suggestions of "proportional" ratings. We therefore converted all ratings to direct percentages rather than doing any normalization or log transforms, which are more typically performed for analyzing ratio data.

### *Performance vs. Subjective Ratings*

Figures 2 and 3 show the performance and subjective rating results across all conditions. To get an overall view of performance for each task, we calculated performance efficiency for each task: percent correct / time in minutes. In essence this is a measure of task completion per unit of time. (See the Common Industry Format for Usability Test Reports, NIST, 1999, for a discussion of this efficiency metric.)

Note the same pattern between the two sets of graphs. Tasks 1 and 4 yielded the highest performance and highest ratings. Likewise, task 2 yielded the lowest performance and lowest ratings.

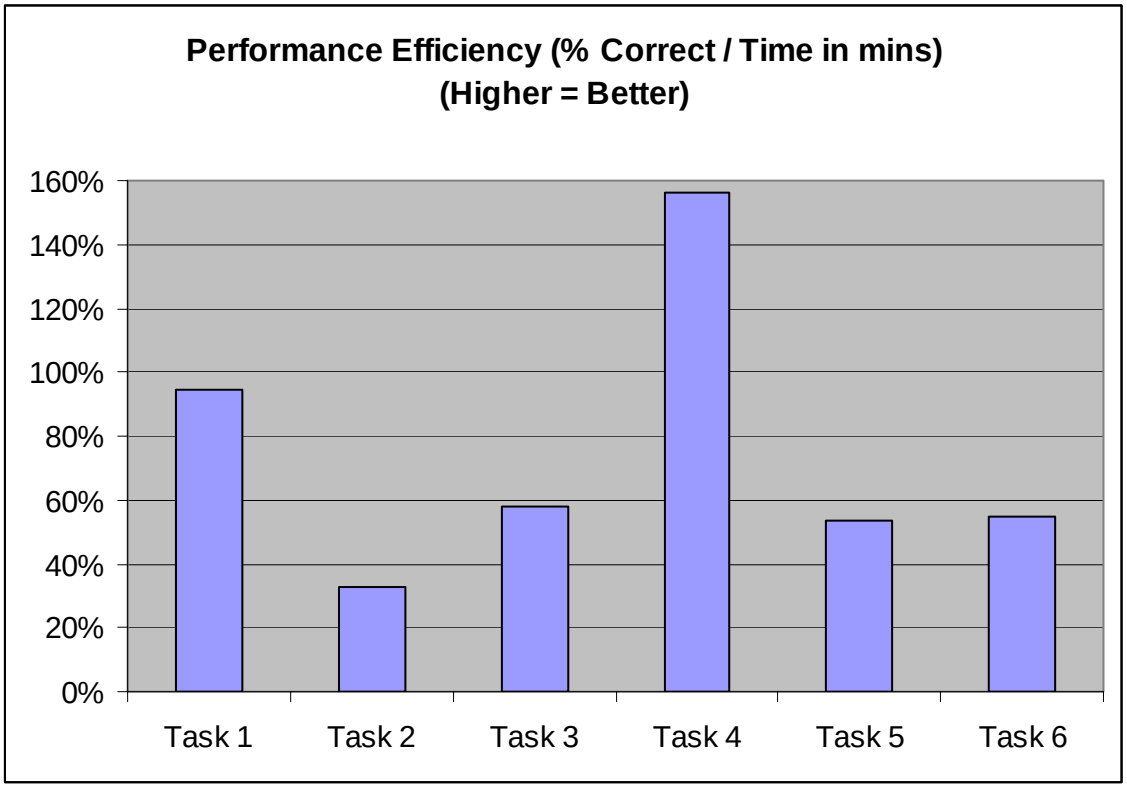


**Figure 2. Performance efficiency for tasks across all conditions.**

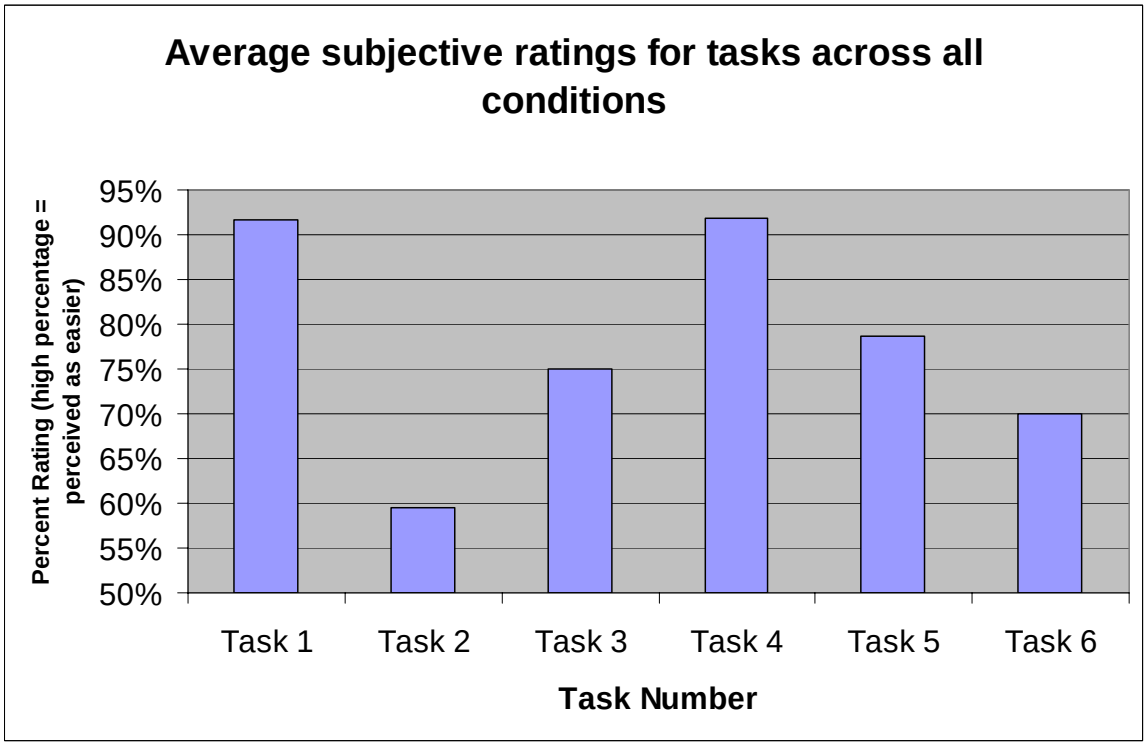


**Figure 3. Average subjective ratings for tasks across all conditions**

In correlating task performance with subjective ratings, correlations were significant for all five conditions ($p < .01$), ranging from a correlation coefficient of $r = .37$ to $r = .46$. The highest correlation was achieved by Condition 4 (the "after" of the before/after expectation ratings). This could possibly be attributed to the fact that users were familiar with all tasks and rating scales before attempting the tasks, thus making ratings more accurate. It could by hypothesized that having rated the expectation of tasks beforehand also adjusts the users' perspectives in a way that becomes more in tune with their actual task experiences. Regardless, simply correlating performance with subjective ratings is not the entire "picture" of what makes a good rating scale. Even if performance and ratings have a 1:1 correlation there is a need to decipher the information to inform us of usability issues necessary to remedy.

As expected, there were no significant differences between conditions for performance efficiency or in the SUS post-study preference ratings.

***Ratings by Condition***

Figure 4 breaks down subjective ratings by condition for the various tasks. The most important implication of this graph is that each of the conditions shows the same pattern of findings with regard to the subjective ratings of the tasks. Note that Condition 4 yielded among the highest ratings for most of the tasks. We again attribute this to an artifact of users having seen and rated their expectations for all tasks before attempting them. Condition 5 yielded the lowest rating for all tasks; this is because it was the only condition transformed as a "true" percentage, i.e., all other conditions could not have a rating below 20% because the 5-point Likert scales were transformed point-by-point to percentages, starting at 20% for a rating of 1. Condition 5, on the other hand, contained lower percentages on average because scores could fall below 20%. We find this difference moot because the important finding is the relative relationships of the different tasks to each other. As long as the same pattern or relationship among tasks is maintained across conditions, as Figure 4 demonstrates, no conclusions can be drawn from this graph about whether one condition may be better or worse than another at detecting differences in tasks. In other words, given our large sample sizes, one would draw the same conclusions about the tasks from all five conditions: Task 2 was clearly the most difficult and should be studied closely to identify any usability issues with the design of the website. To a lesser extent, Task 6 was also difficult for the participants.

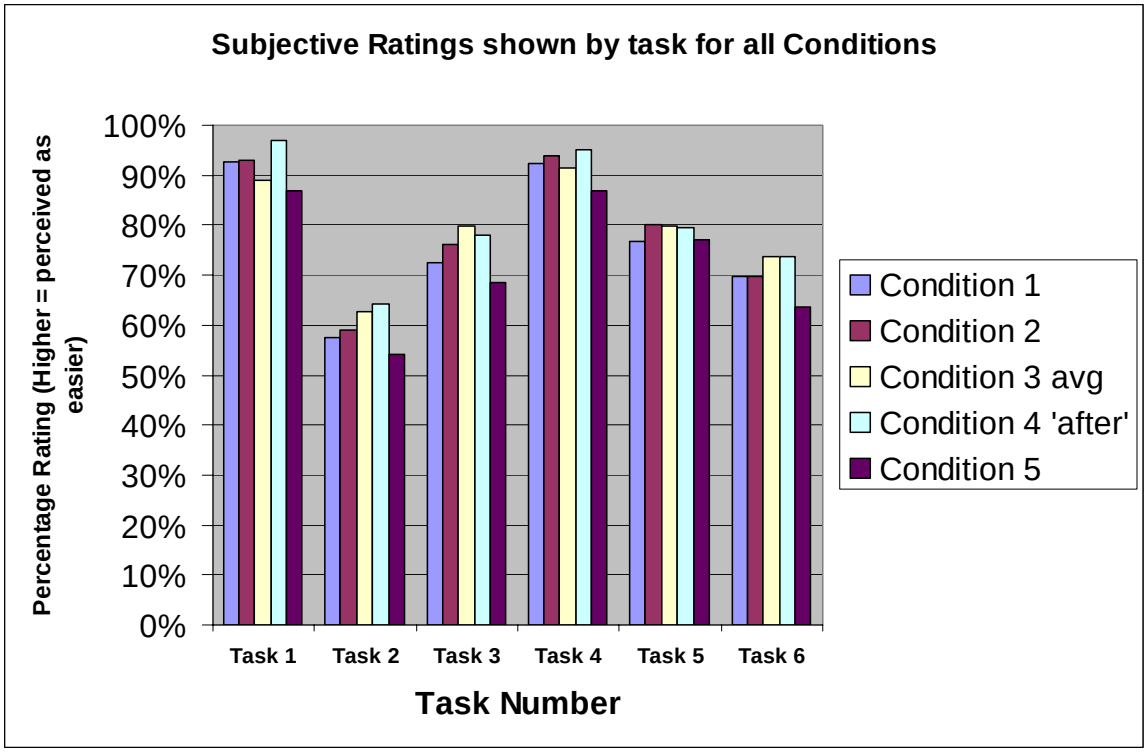


**Figure 4. Subjective Ratings shown by task for all Conditions**

## Sub-sampling Analysis

One of the goals of this study was to determine if one approach to post-task ratings yields more reliable results than the others at the smaller sample sizes typical of most lab-based usability tests. To address that question, we conducted a sub-sampling analysis similar to one used by Tullis & Stetson (2004). To simplify the process, we wrote a program that took a large number of random samples of different sizes from our complete dataset for each condition. Specifically, the program took 1,000 random samples from the full dataset at sub-sample sizes ranging from 3 to 29 in increments of 2. For each random sample, the correlation between the average ratings for the six tasks and the average ratings for those six tasks in the full dataset was calculated. The results are shown in Figure 5.

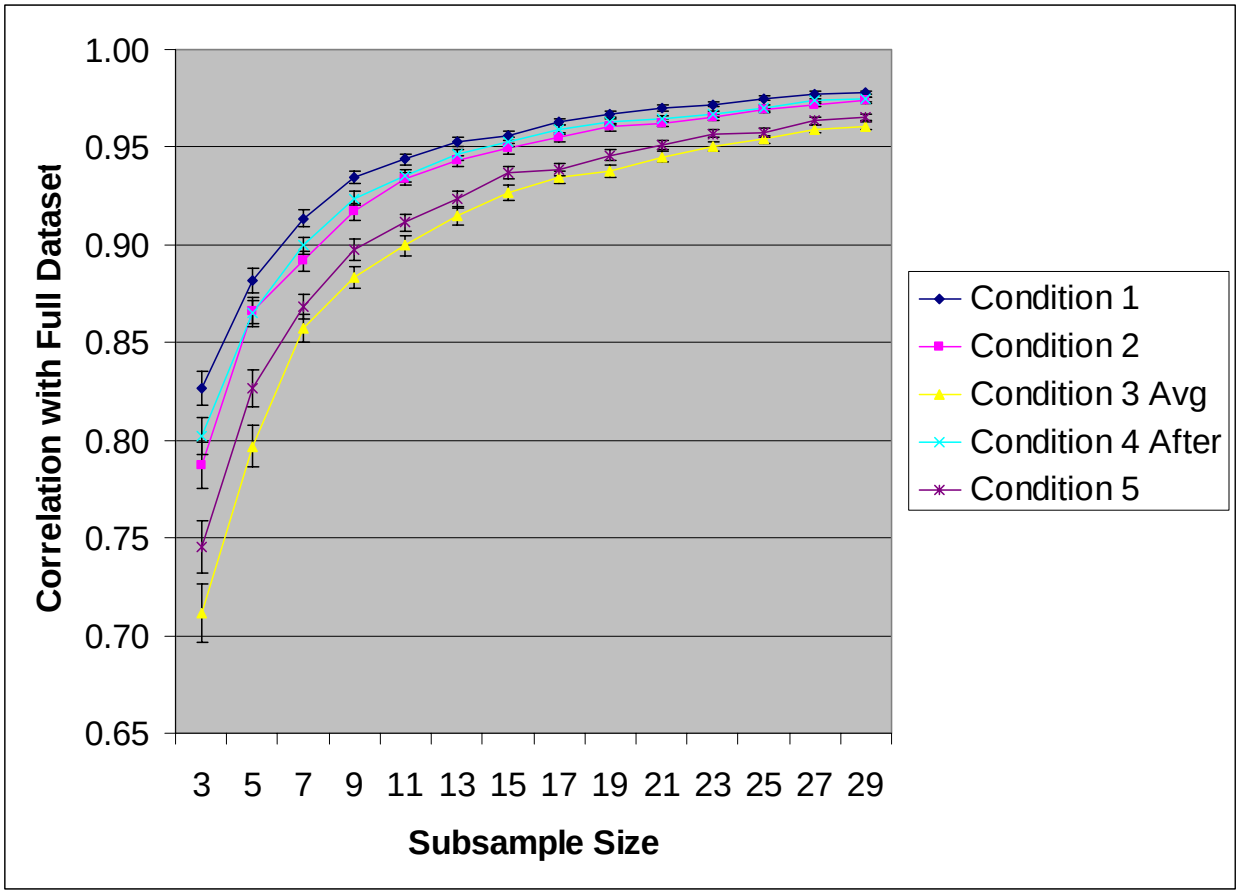


**Figure 5. Average correlations between ratings for the six tasks in sub-samples of various sizes and the full dataset, for each condition. Error bars represent the 95% confidence interval for the mean.**

As can be seen from Figure 5, all of the conditions yielded good results at the larger sample sizes, with all of them resulting in correlations of at least r = .95 at sample sizes of 23 or more. But at the smaller sample sizes typical of most usability tests, there were significant differences. At these smaller sizes, Condition 1 outperformed all the others, consistently yielding slightly higher correlations with the full dataset. Basically this means that participants were slightly less variable in their responses to this rating scale.

## Conclusions

The following conclusions appear warranted from this study:

- All five of the post-task rating techniques yielded significant correlations with the performance data for these tasks. This means that any of them could be reasonable "surrogates" for identifying tasks that users had difficulty with in situations where it may not be practical to collect performance data.
- All five post-task rating techniques yielded significant differences among the ratings of the tasks, given our large sample size. This indicates that at least with large sample sizes any of them could be used to identify which tasks to focus on when trying to identify usability issues.
- The sub-sampling analysis indicated that perhaps the simplest post-task rating scale, Condition 1 ("Overall this task was: Very Easy … Very Difficult"), yielded the most consistent results at smaller sample sizes. At a sample size of 7, which is typical of many usability tests, its correlation was .91, which was significantly higher than the others.

Note that this study focused on identifying which tasks were perceived as the most difficult, in comparison to the other tasks. The assumption was that these tasks, and the features of the website that support them, would then be candidates for closer inspection to identify usability issues. On the other hand, information of a different type could be available from a comparison of the "before" and "after" ratings in Condition 4. For example, one might decide that the primary tasks to focus on are the ones that have the greatest *negative change* from the users' expected difficulty to their actual difficulty, regardless of how the actual difficulty compared to the other tasks.

## References

Albert, W., and Dixon, E. (2003). Is This What You Expected? The Use of Expectation Measures in Usability Testing. Proceedings of Usability Professionals Association 2003 Conference, Scottsdale, AZ, June 2003.

Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. SIGCHI Bulletin, 23, 1, 78-81. Also see http://www.acm.org/~perlman/question.cgi?form=ASQ.

McGee, M. (2004). Master Usability Scaling: Magnitude Estimation and Master Scaling Applied to Usability Measurement. Proceedings of CHI 2004 Conference on Human Factors in Computer Systems, Vienna, Austria, pp. 335-342. New York, NY: ACM.

NIST (1999). Common Industry Format for Usability Test Reports, version 1.1. Retrieved from http://zing.ncsl.nist.gov/iusr/documents/cifv1.1b.htm on April 18, 2006.

Tullis, T. and Stetson, J. (2004). A Comparison of Questionnaires for Assessing Website Usability, Usability Professionals Association (UPA) 2004 Conference, Minneapolis, MN, June 7-11, 2004.