EXPECTED USABILITY MAGNITUDE ESTIMATION

Aaron Rich
Oracle Corporation
Redwood Shores, CA, USA

Mick McGee, Ph.D.
Oracle Corporation
Redwood Shores, CA, USA

Usability measures typically focus on actual user experiences while largely ignoring the impact of user expectations. User expectations provide insight into overall usability, user satisfaction, and priority of usability problems. Beyond test results, communicating user expectations can offset the negative connotation many development teams have of usability by showing examples where expectations are exceeded. This paper describes the expected Usability Magnitude Estimation (UME) method to assess user expectations in usability tests. The method is more valid, robust, and theoretically based than existing methods. It allows measurement of expectations that is easy to administer, simple to analyze, and provides actual and expected usability ratings along the same ratio scale of usability. Expectation data is used to classify tasks into empirically derived design strategy groupings based on refined theory. Overall, the method positively contributes to usability results and development team relationships.

INTRODUCTION

Usability practitioners are mainly concerned with assessing a user's actual experience with an interface to determine its usability. This is typically accomplished through the use of various measures that assess the key components of usability generally held in the field to be efficiency, effectiveness, and satisfaction (ISO, 1998).

Satisfaction is typically measured with post-test questionnaires and Likert-style ratings. These measures attempt to judge the level of satisfaction in a user's actual experience with a software interface. However, users' expectations are an important element of satisfaction that often gets overlooked. Expectations of usability play an important role in how users perceive their actual usability experience.

The implications of user expectations are described in the expectancy disconfirmation theory (Oliver, 1977). It states if quality is below expectations, dissatisfaction occurs. Conversely, if quality is above expectations, satisfaction occurs.

For example (in a usability context), a user expecting an easy task that turns out to be difficult would be less satisfied than a user expecting a difficult task that is indeed difficult. Typical usability measures would show these two tasks were difficult to complete, took the same time, etc., but would not provide any insight into the difference in satisfaction. Usability expectation measures can provide this insight. A sample benefit would be assigning higher priority to cases where usability problems caused tasks to be more difficult than expected.

Usability Expectation Measures

Albert and Dixon (2003) proposed comparing expectation ratings (the expected usability of a task) with experience ratings (the actual perceived usability of a task) using anchored Likert scales. Their method involved participants reviewing the task list prior to testing and assigning expected usability ratings for all tasks. Then, after each task was completed, an experience usability rating was assigned using the same Likert scale. Actual and expected scores for each task were then analyzed using a wide range of statistical procedures.

For advising design strategy, Albert and Dixon (2003) defined four distinct groups of tasks based on actual versus expected ratings, shown in Table 1. "Fix it fast" tasks were expected to be easy, but were actually difficult. "Promote it" tasks were expected to be difficult, but were actually easy. "Don't touch it" tasks were expected to be easy and were actually easy. "Big opportunity" tasks were expected to be difficult and were difficult to complete.

Table 1. Expected vs. Actual Usability Task Categorization.

		Expected Usability	
		Good	Bad
Actual Usability	Good	Don't touch it	Promote it
	Bad	Fix it fast	Big opportunity

Measuring usability expectations is a great concept, as are the intuitive task classifications; however, the Albert and Dixon (2003) method has a few key drawbacks. First, Likert scales are known to be problematic for implying underlying continuums, which lead to empirical analyses of questionable validity (Badia and Runyon, 1982).

Second, possibly caused by the limitations of the Likert scale analysis, Albert and Dixon do not provide a specific method for determining which tasks belong to which categories. They propose a wide range of methods: statistical (descriptive statistics, ANOVAs, correlation, and weighted usability factors), various graphically defined regions on scatter plots, and simple visual inspections.

Lastly, related to both the previous issues, Albert and Dixon recommend 12 participants per usability test to find statistical differences with the Likert scales. The reliance on Likert scales, the confusion over task classification, and the requirement for many users led us to seek a more valid, easier to interpret, and practical alternative.

Usability Magnitude Estimation

Usability Magnitude Estimation (UME) is a measure that can be adapted to assess expectations without inheriting the aforementioned limitations. UME is a subjective assessment method where participants assign usability values to targets using ratio-based number assignment (McGee, 2003). Usability ratings are made based on the following objective definition of usability:

> "Usability is your perception of how consistent, efficient, productive, organized, easy to use, intuitive, and straightforward it is to accomplish tasks within a system"

UME is quick and easy to administer, has no upper or lower limits to ratings, and produces data that is appropriate for statistical analysis (the geometric data reduction process is described in McGee, 2003). Meaningful data-driven statements of magnitude (i.e., task A is twice as usable as task B) can be made since it is based on a ratio scale.

UME can be adapted to assess expectation measures by collecting estimates of expected usability and comparing them to estimates of actual usability experienced. A key benefit is that the same underlying usability scale is ensured for both the actual and expected ratings, allowing valid comparisons along a true continuum. This paper describes how to implement expected UME, analyze and interpret results, and communicate findings to development teams.

METHOD

A usability test of a prototype Business Intelligence application was conducted with six participants at the Oracle usability labs. The application allowed project portfolio analysts to identify, analyze, prioritize, and select optimal project investments for given business scenarios. Participants completed ten tasks, providing expected and actual usability magnitude estimates for each task. The ten tasks tested were:

1. Retrieving basic information on a project
2. Retrieving project resourcing information
3. Ranking projects
4. Saving changes in a scenario as a new scenario
5. Delaying a project in a portfolio
6. Editing project information
7. Retrieving project scorecard information
8. Retrieving executive summary metrics
9. Switching portfolio scenarios
10. Adding new projects to a portfolio

Expected Usability Magnitude Estimation Procedure

To collect expected UME scores, participants were first instructed in UME measurement. Participants then performed a practice task prior to the start of the usability test to familiarize themselves with how magnitude estimation is used. Then, prior to the start of the usability test, users were instructed that they would be making two UME measures. The first measurement was based on the user's expectation of a tasks' usability, made after reading the instructions for a task, but before starting the task. Immediately following the completion of a task, users made a second rating of usability based on their actual experience in performing the task.

RESULTS

The expected and actual average scores for each task are shown in Table 2. Expected UME scores ranged from 17.55 to 24.29 while the actual UME scores ranged from 15.75 to 28.87. The average absolute value percent difference from expected to actual UME across all tasks was over 15%. The positive and negative extremes were 30.6% and -33.2%.

Table 2. Expected, Actual, and % Difference UME Scores.

Task	Expected UME	Actual UME	% Difference
1	20.60	23.62	14.6%
2	20.54	18.24	-11.6%
3	17.55	17.25	-1.7%
4	20.28	20.90	3.1%
5	19.40	24.53	26.4%
6	24.29	22.89	-5.8%
7	20.07	21.96	9.4%
8	23.59	15.75	-33.2%
9	19.17	23.90	24.7%
10	22.10	28.87	30.6%

Design strategy groupings of tasks were empirically determined based on our theoretical depiction of actual to expected usability, shown in Figure 1. In essence, the plot is a cost/benefit analysis of usability expectation disconfirmation.

The basis of the cost/benefit analysis is where expectations meet experience, defined as "Opportunity". Decreasing values along this fixed continuum have more room to improve in actual usability, and less expectations to meet, thereby increasing the area available to increase satisfaction due to positive expectation disconfirmation (i.e., "Big opportunity"). Increasing values have less area of potential positive expectation disconfirmation. With expectations and actual usability already high, the costs are too prohibitive to obtain tangible gains (i.e., "Don't touch it").

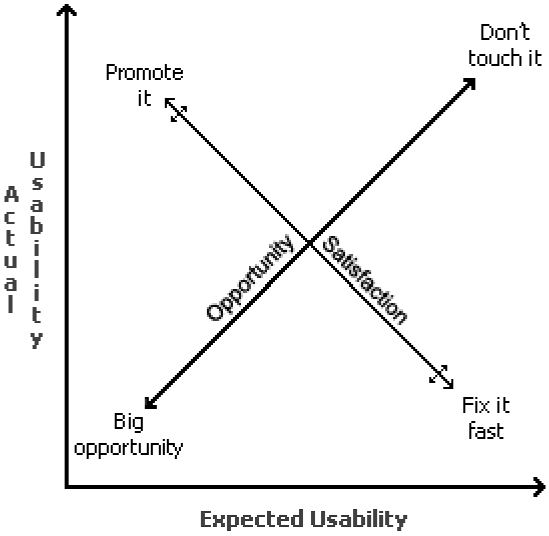

Figure 1. Theoretical Usability Expectation Disconfirmation.

The Opportunity line divides the expectation disconfirmation plane. Any continuum orthogonal to where usability experience meets expectations is an instantiation of expectation disconfirmation, defined as "Satisfaction". Increasing values along this continuum (i.e., lower expected usability and higher actual usability) lead to ever-greater positive disconfirmation and satisfaction (i.e., "Promote it"). Decreasing values (i.e., higher expected usability and lower actual usability) lead to ever-greater negative disconfirmation and dissatisfaction (i.e., "Fix it fast").

Any given data point is considered by: 1) its actual expectation disconfirmation (distance from the Opportunity line, along an orthogonal Satisfaction vector), 2) it's potential to improve expectation disconfirmation (area above or below the fixed Opportunity line), 3) the relative costs (e.g. effort) for improving actual usability, and 4) the probable regression to the mean over time (i.e., actual usability) for expectations.

The first two data point considerations can be represented by percentage differences from the Opportunity and Satisfaction continuums. To guide these thresholds, Weber's Law says a 10% change in stimulus intensity is necessary for a just noticeable difference (Gescheider, 1997). Controlled experiments using UME corroborate this law, showing minimum significant UME differences between 5% and 15% (McGee, 2003). For this paper, we chose the 15% conservative minimum threshold to categorize expectation usability data. "Just" noticeable differences are not sufficient for discrete categorization.

Figure 2 shows the Business Intelligence data for each task plotted in the actual vs. expected format of Figure 1. The 15% categorical bands can be overlaid on this plot; however, it is easier to view the categories when the plot is rotated around the axes of Opportunity and Satisfaction and shown in percent difference from those lines respectively, shown in Figure 3.

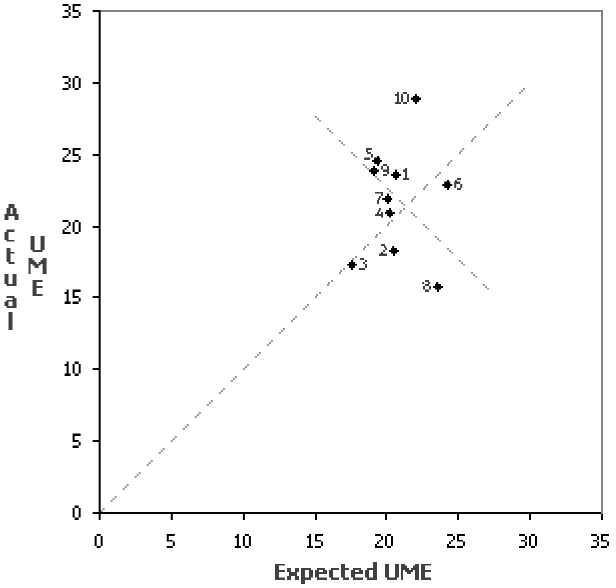

Figure 2. Business Intelligence Data By Task: Actual vs. Expected UME.

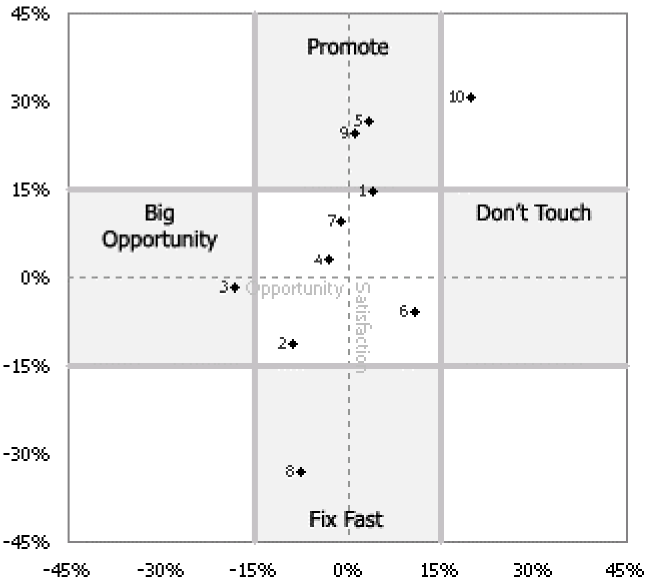

Figure 3. Business Intelligence Data By Task: Percent Difference from Opportunity and Satisfaction.

Calculations for percent difference from Opportunity for each data point are straightforward:

(Actual UME - Expected UME) / Expected UME * 100%

E.g., for task 8:

(15.75 - 23.59) / 23.59 * 100% = -33.23%

Calculations for percent difference from Satisfaction first need to determine known points with the given data.

The only known point along the Satisfaction line is where expectations meet actual usability. This is the overall average of all the actual and expected usability scores; e.g., 21.28 in the test described (x=21.28 y=21.28). Then, each data point's position relative to this known point needs to be determined; by finding where each individual data point's expectations meet actual usability. This is simply the average of the actual and expected usability for each data point; e.g., 19.67 for task 8 (x=19.67 y=19.67). Thus, the formula for percent difference from Satisfaction for each data point (*i*) is:

(Ave. UME $_i$ - Ave. UME All) / Ave. UME All * 100%

E.g., for task 8:

(19.67 - 21.28) / 21.28 *100% = -7.54%

Two tasks (delay projects and switch scenarios tasks) were classified into "Promote it". One task (executive summary task) was grouped into "Fix it fast". One task (rank projects task) met the "Big opportunity" criteria. No tasks fell into the "Don't touch it" category.

DISCUSSION

Expected UME proved to be a useful method for evaluating the impact of user's expectations on perceived usability. The method contained a number of positive attributes that were observed through its use in the described usability test.

Implementation and UME Analysis

Expected UME was simple to implement and easy for participants to understand. Participants were required to make only one additional rating compared to the conventional UME methodology.

One concern was whether expectation scores would provide enough discrimination among multiple tasks and between the corresponding experience scores. The rankings showed that participants did perceive real differences between expectations and actual experience as well as differences in expectations across tasks.

Another concern was order effects. Expectations change over time, particularly with a prototype where users have no explicit prior experience. No order effects were found with the described study; however, order effects should be considered in the experimental design of future studies. Where possible, randomization of task order could alleviate order effects. Or, perhaps more appropriately, expectations should simply be allowed to adapt naturally, even within the course of a single study.

Another benefit of using magnitude estimation as the basis for expectation ratings is its readiness for use in parametric statistical analysis. Additional analyses could be conducted if desired; assuming associated experimental controls were used.

Design Strategy Grouping Theory

Albert and Dixon (2003) described significant flexibility in determining how to group tasks into each of the four previously described design strategy groupings. The problem centered on the subjective nature of creating groupings.

Expected UME has the advantage of being based on the cost/benefit analysis of the usability expectation disconfirmation theory. This allows design strategy groupings to be determined empirically. This makes the results explicit and defensible, which is important when presenting to development teams and getting commitment to usability recommendations.

Prioritizing Usability Issues

The main benefit of using expectation measures in usability is to help prioritize the usability issues. The tasks grouped in the "Fix it fast" category are the tasks with usability issues that need to be addressed first. These tasks have unexpected poor usability. The "executive summary" task in this usability test was an example of this. Participants expected to easily find an executive summary metric in the application. They were surprised when they had difficulty finding it since it was located in a difficult to find page of the application that was not intuitive to most users. The case can be made to development teams that addressing the usability problems in these types of tasks will have the greatest positive impact on the product.

The tasks falling in the "Promote it" grouping provide development teams with positive reinforcement and a marketing opportunity. The "switch scenarios" task was an example of this. Participants thought it would be difficult to switch between various scenarios to compare project costs. They were pleased to discover that a control was included that allowed switching between scenarios to instantly update project data with a single click. The resultant usability was much higher than initially expected. Usability professionals often have a reputation for providing "bad news" about the usability of a product. Tasks like this that significantly exceed user expectations can be used to show what has been well designed.

Another group of tasks that should receive high priority for fixing usability issues is the "Big opportunity" grouping. For example, the "rank projects" task in this usability test was expected to be difficult and have poorer usability compared to other tasks. Participants thought this task would be performed using difficult to find information and features. This was in fact the case and caused actual usability ratings to be low as well. Focusing on improving the usability of tasks like this that fall into the "Big opportunity" group can elevate them to the "Promote it" grouping in future versions.

Non-Categories. The four design strategy groupings are a useful way to present priorities of tasks to development teams. However, they are not intended to be absolute discrete categories. As the theory plot shows, the categories are actually made from continuums.

There are four "corners" not labeled in the percentage plot, along with the "normal" center square. The last two considerations for any given data point (relative costs for improving usability and regression to the mean for expectations over time) are the subjective differentiators for points falling outside of the labeled categories.

Tasks falling between the "Fix it fast" and "Big opportunity" groupings are still important to fix because of their overall poor actual usability. However, expectations are not so high that they are "Fix it fast" priority items, nor are expectations so low that they are "Big opportunity" priorities. Depending on the actual position of the data point, the extent that expectations might regress to actual experience can weigh on the importance of prioritizing these issues.

Tasks between "Promote it" and "Don't touch it" also have mixed expectations; however, actual usability is high. The marketing value is lessened since expectations are already average for this product, and expected to trend higher over time to meet actual usability experience. Task 10 (adding new projects to a portfolio) is an example of a task that fell into this "corner" zone. Adding a new project was expected by participants to be relatively easy since it was perceived to be a common core task that the application was designed to support. As it turned out, the application was so well designed for this particular type of task that actual usability was scored higher than the expectations, even though the expectations were high to begin with. It would be difficult and costly to improve the usability of this task enough to place it in the "Promote it" category, especially since the expectations will most likely increase in the future as well.

Tasks between "Promote it" and "Big opportunity" have average usability and low expectations. Tasks between "Don't touch it" and "Fix it fast" have average usability and high expectations. Cost/benefit associated with improving actual usability for these data points may have more importance than trending expectations. Improving the usability of a task with high expectations and average actual usability might not be worth the effort.

In terms of the expectation disconfirmation analysis, the "normal" center square is comprised of the lower priority usability problems. All tasks, specifically the center tasks, need to consider the entirety of the usability information collected within an activity (both qualitative and quantitative) to refine priority.

Relative Expectations

This paper is primarily related to expectations for a given set of data. However, the theory can be extended to comparisons over time and between products. For example, one product could have much lower actual usability than another, but be more satisfying because expectations are even lower.

On the theoretical graph shown in Figure 1, while the Opportunity line is always fixed (where expectations meet experience), the Satisfaction line is fixed only for given data. Considering the current study, a corresponding mobile Business Intelligence prototype would likely have lower actual usability scores and lower expectations, which would move the Satisfaction line much closer to the origin (0,0). If expectations are low enough, mobile users might be more satisfied than desktop users. New technology (e.g. mobile) is expected to be hard to use, but users are usually happy to have the new functionality regardless of usability. As expectations increase over time for new technology, actual usability will have to improve to maintain relative high satisfaction.

CONCLUSIONS

User expectations are a valuable addition to usability. The insight into satisfaction allows more impactful usability conclusions to be made. Expected UME has a number of positive characteristics for measuring user expectations:

- Easy to administer and analyze
- Uses the same underlying ratio scale for actual and expected usability ratings, allowing valid comparisons
- Provides a theory-based and empirical usability issue prioritization strategy

REFERENCES

Albert, W.S. and Dixon, E. (2003). Is this what you expected? The use of expectation measures in usability testing. *Proc. Usability Professionals Association, 12thAnnual Conference*, 10th paper.

Badia, P. and Runyon, R.P. (1982). *Fundamentals of behavioral research.* Random House.

Gescheider, G.A. (1997). *Psychophysics: The Fundamentals, 3rd Ed.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

ISO 9241-11:1998(E). (1998-03-15). *Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on Usability*. International Organization for Standardization, Switzerland.

McGee, M. (2003). Usability magnitude estimation. *Proc. Human Factors and Ergonomics Society, 47th Annual Meeting*, (691-695).

Oliver, R.L. (1977). Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology*, 62(4), (480-486).