Interactions (2006) 13(6): 42-43

Practical Issues in Usability Measurement

Nigel Bevan □ Professional Usability Services □ nigel@nigelbevan.com

How do you select the most useful and appropriate usability measures?

If the purpose of the test is to decide whether a product has adequate usability (the type of test the CIF is intended for), measures should focus on the end result (as defined in ISO 9241-11):

- Effectiveness: "accuracy and completeness." Error-free completion of tasks is important in both business and consumer applications.
- Efficiency: "resources expended." How quickly a user can perform work is critical for business productivity.
- Satisfaction: "positive attitudes toward the use of the product." Satisfaction is a success factor for any products with discretionary use; it's essential for maintaining workforce motivation.

This is a black-box view of usability: *what* is achieved, rather than how. The CIF requires as a minimum measures of the percentage of participants achieving each task goal, the mean time taken to complete each task, and a measure of user satisfaction.

Other common usability measures such as frequency of errors, frequency of assists, and frequency of accesses to help or documentation all tell you something about why a product is or is not usable, so these measures can be important symptoms of particular usability defects. But they are less important from a user and business perspective.

Additional Summative Measures

While downplaying the importance of many common formative measures, the CIF and CISU-R suggest some additional summative measures.

Partial goal achievement. In some cases goals may be only partially achieved, producing useful but suboptimal results (for example, for a travel Web site, not finding the cheapest air ticket). In a formative test, partial success is often judged by counting the number of assists to obtain a correct result. But for summative testing, the task is complete when the user declares that they are finished. Suboptimal results can then be scored in terms of requisite components (e.g., up to 50 percent for the correct date and time and up to 50 percent for the best fare). Overall effectiveness would be the mean value across participants.

Relative user efficiency. As the time taken will depend on the nature of the task, it is possible to make direct comparisons of efficiency data only when the same task is performed. A useful measure that is independent of the task is relative user efficiency: how long a user takes in comparison with an expert. This is measured as the mean time taken by users who successfully achieve a goal divided by the time taken by an expert. It highlights the potential usability gap between typical users and an expert user (often showing that it takes normal users two or three times longer to complete a task than an expert).

Productivity. Another measure of efficiency that is sometimes used is completion rate divided by task time, which gives a classical measure of productivity: the rate at which goods or services are produced. It provides a measure that trades off time against accuracy and completeness, but the units (percent task completion per minute) can make it difficult to interpret.

Interactions (2006) 13(6): 42-43

Satisfaction Questionnaires

Satisfaction measures are most reliable when a psychometrically validated questionnaire is used [1].

Satisfaction questionnaires measure users' expectations, which means that it is possible to use a questionnaire to make comparisons across products. For some questionnaires, data is available to compare the results with industry norms. This makes it easier to interpret whether an individual result is good or bad.

Measuring in different contexts

Resource constraints mean that usability testing is typically carried out in the most common context of use. But many important usability issues arise only in less-common contexts:

- *Learnability*: usability measured for the task of achieving adequate performance, for example, by completing a training course or through the use of learning materials.
- *Accessibility*: usability measured for users with particular disabilities.
- *Universality*: usability measured in a range of different contexts and cultures.
- *Risk*: usability measured in situations that may have business or personal risk.

Even if there are limited resources for usability testing in these situations, it is useful to establish targets for effectiveness, efficiency, and satisfaction, so that they are accounted for in the design.

Interpreting the results.

Results should be reported with a confidence interval [2]. If there is no overlap in the confidence intervals when making comparisons, you can be sure there is a statistically significant difference.

The main risk in interpreting statistical data, particularly from small samples, is that the results can be generalized only to the population from which the sample of participants is taken. Although demographics are a useful starting point for selecting participants, the participants also need to be representative of the range of individual characteristics that may influence usability; otherwise the results may be misleading.

Sample Size

The sample size required to achieve specific confidence intervals can be estimated if the variance of the dependent measures is known. But this will be the case only where there is a past history of similar measures. Otherwise one can make only broad generalizations. For task time and satisfaction data, experience has shown that a minimum of eight to ten participants is usually required to begin to make reliable estimates (the CIF requires a minimum of eight). However, if there are major usability problems, even the results from testing three to five participants would be likely to provide advance warning of a potential problem (for example, if none of the participants can complete their tasks, or tasks times are twice as long as expected).

Reliable estimates of success rate require much larger samples for results coded just as success or failure. The binomial distribution shows, for example, that to achieve 95 percent confidence of an 80 percent success rate, all participants in a sample of 14 would have to succeed, or 21 out of a sample of 22.

Interactions (2006) 13(6): 42-43

Conclusions

Although the most popular form of usability testing is formative—to get quick feedback to improve the usability of products using small numbers of participants—it leaves open the risk of inadequate final usability. Using the CISU-R to establish usability requirements that can be tested and reported in the CIF can provide the foundation for a mature approach to managing usability in the development process.

References

1. Kirakowski, J. 2000 Questionnaires in usability engineering. www.ucc.ie/hfrg/resources/qfaq1.html (accessed 1 Jul 2006)
2. Sauro, J, 2005. Confidence interval for task times in usability tests. http://www.measuringusability.com/time_intervals.php (accessed 1 July 2006).