

Designing a User-Based Evaluation

Nigel Bevan

June 2007

It is important to decide the high-level objectives for conducting a usability test. For example:

- Is the priority to identify usability problems or to validate usability?
- Is it exploratory or being used to compare design alternatives?

Usability testing should be carried out carefully and systematically, otherwise the resulting data may not be valid or reliable. O'Hara et al (2005) describe the need for the following types of validity:

- *External validity*: extent to which the context of use for the test is realistic.
- *Construct validity*: extent to which the measures are representative of user performance and satisfaction.
- *Internal validity*: extent to which the test is properly designed.
- *Statistical validity*: extent to which the statistical conclusions are valid.

Poor data may lead to poor design decisions and ultimately an error prone and unsafe product. Therefore, usability testing protocols should be developed in collaboration with professional usability specialists. Non-specialists under the direction and training of professionals can handle the actual execution and reporting of the testing.

Design for the Expected Context of Use

Select the most important tasks and user groups to be tested (e.g. the most frequent or the most critical). It is important that the users, tasks and environment used for the test are representative of the intended context of use.

Other contexts

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

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

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 supported by the design.

Select Test Tasks Related to Goals of Use

Information about the goals for which the product will be used and the intended context of use should be obtained from the relevant users and stakeholders. Decide which goals and tasks are to be tested. Usually, this will be the most frequent and/or critical goals

that the product is intended to support, and the most frequent and/or critical tasks expected to be performed using the product for the goals.

The goals to be tested should be expressed in terms of the intended outcome of the task activity expressed independently of the means by which it is achieved. The criteria for complete and accurate task achievement should be specified.

For each goal and task, the condition(s) associated with successful goal attainment should be identified together with a way of monitoring when it has been achieved. Because successful completion is the criterion, it is generally appropriate to include alternative methods of achieving the goal(s) as positive results, unless the means used have undesirable consequences, such as posing a risk to health or safety, or causing damage to the product.

A maximum amount of time allowed for successful goal achievement should be established.

If there is more than one main goal to be tested it will be necessary to decide the order in which the tasks are to be undertaken. In general, if there is a normal sequence in which tasks would be undertaken this should be the order for testing. If no normal sequence exists it will be necessary to avoid testing in one fixed sequence, in order to avoid arbitrary order effects. In this case, the order may be systematically varied or assigned randomly for each person tested.

If there is any uncertainty over how users will actually use the product, it is desirable to allow users to identify some of their own goals, as this will clarify user requirements and may explore aspects of the product that would not be encountered by pre-defined tasks.

Select Measures

For a summative test, the usual measures will be:

- *Effectiveness*: whether (or to what extent) the task is completed successfully.
- *Efficiency*: The time taken to achieve the goal(s).
- *Satisfaction*: Measured by a questionnaire administered after completion or termination of all the tasks. Satisfaction measures are most reliable when a psychometrically validated questionnaire (such as SUS (Brooke, 1996)) is used (Kirakowski, 2000). Other questionnaires such as TLX (Hart & Staveland, 1988) for workload may also be appropriate.

Recruit a Representative Sample of Users

Usability test participants should be representative of the intended users. First decide whether the participants should be representative of the whole user population or of one or more identified subgroups. Identify which characteristics of the users may influence the usability of the product. If the intended user group includes people such as the elderly or disabled, decide whether these are to be included in the test group, or tested as a separate group. A sample representing the intended users should be recruited to take part in the test.

Previous experience or training in the use of this product or similar products or tasks is often a key factor. Selecting participants with no previous experience can simulate the worst case situation, but if, in reality, users would have appropriate experience or training, it should be provided before the evaluation.

Sample size

For formative testing, five to eight participants are often suggested as sufficient to identify the main usability problems. Five to eight would be sufficient to identify 80% of problems if the probability of each problem being found by one user is between 18% and 30% (Sauro, 2006). When the probability of finding a problem is lower (for example if different users take different routes through the software), much larger numbers would be required (Lewis, 2006; Perfetti & Landesman, 2001).

In general several iterations of testing with a few participants will be more effective than only one test with many participants.

For summative testing, representative sampling should be used to set up a sample that models the distribution of relevant user characteristics within the intended user population. The number of participants required to obtain reliable results will depend on the diversity of the user characteristics, and what are acceptable confidence intervals to obtain in the overall results. Measures such as task time and satisfaction scores produce continuous data, and, for well-defined user groups, useful results can be obtained with as few as eight participants (ISO, 2006). The actual number required can be calculated if you know the variance in the data (from a previous similar study or a pilot study) and the confidence level required (Lewis, 2006).

However if there are major usability problems, even the results from testing 3-5 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).

Larger samples are required to obtain reliable estimates for binary data such as successful completion rate. The binomial distribution shows, for example, that to achieve 95% confidence of an 80% success rate, all participants in a sample of 14 would have to succeed, or 21 out of a sample of 22.

This discrepancy between the numbers required for different types of data mean that in some cases there will be insufficient participants to make an accurate prediction of success rate.

Comparisons

Test results can be compared with a predefined requirement, or with other test results. Comparisons with other test results are only meaningful if the tests were carried out in the same context of use. The results should include a statement of the statistical probability that any differences may be due to chance.

Test Environment

Testing should if possible be carried out in the normal environment in which the product would be used. If a usability laboratory is used, aspects of the normal conditions of use that might influence usability should be recreated.

Planning

Plan Test Procedure

If the main purpose of the test is summative, to obtain reliable measures it is important that the test procedure is as natural as possible, with only forms of assistance that would be available in the real world. The participant should carry out the task alone and not be asked to think aloud. As interaction with the moderator should be minimized, it is preferable for the moderator not to be present in the same room.

If the main purpose of the test is to identify usability problems, there are several options, depending on whether some usability measures are also required:

- Ask the participant to think aloud (approximately one third of people can do this naturally, one third need encouragement, and one third have difficulty). Although this changes behavior, reasonable estimates of task time and success rate can still be obtained.
- Probe the participant to explain why they are making particular choices, and what they would expect to happen next. This will affect task time, but estimates of success rate can still be obtained.
- If (and only if) the participant is unable to continue, give hints. This will enable feedback to be obtained on subsequent parts of the task, but the trial should be counted as a failure.

Other activities include:

- Produce a task scenario and input data and write instructions for the user (tell the user what to achieve, not how to do it).
- Plan sessions allowing time for giving instructions, running the test, answering a questionnaire, and a post-test interview.
- Produce a list of questions to be asked in the post test questionnaire (determined by the purpose of the test and the concerns of the stakeholders).
- Invite developers to observe the sessions if possible. An alternative is to videotape the sessions and show developers edited clips of the usability problems. This type of “shared representation” is very effective in communicating the significance of the results.
- Two administrators are normally required to share the activities of instructing and interviewing the user, operating video equipment (if used), noting problems, and speaking to any observers.

A written script should be prepared for each task goal, containing instructions that are read to each user. The script should include a description of the scenario within which testing is taking place and should state the particular goal and the conditions which apply to it. No hints should be included on how to achieve the goal or which features to use.

If measures are being taken, those conducting the test should note the time taken to achieve each goal. If the user reaches the maximum amount of time allowed without attaining the goal, they should be asked to stop, and, if appropriate, move on to the next goal.

Use of a Usability Laboratory and Video Recording

Usability tests are frequently carried out in a usability laboratory where the participant and observers are separated by a one-way mirror, and the participant and screen are monitored on video and video recorded. This enables the observers from the development team to talk among themselves while the participant works undisturbed.

The video-recording can be used to check back on any actions not fully understood in real time, but the major benefit is to be able to extract video clips to show to people who are unable to watch live trials. Marking up events and retrieving video segments can be simplified by the use of video analysis tools such as Morae™.

Eye-tracking equipment can be used to create a video playback of the screen with the user's eye scan traces superimposed. This can be particularly useful to help understand why users do not notice particular screen elements.

In the absence of a lab, portable equipment can be used to set up a video link between two rooms. If video is not essential, testing can be carried out in one room, but preferably without additional observers.

Results

Produce a list of usability problems, categorized by importance (post-it-notes can be used to sort the problems), and an overview of the types of problems encountered. Care should be taken in over-interpreting results from a single participant, although if they reveal an obvious flaw in a prototype it makes sense to fix this immediately (Medlock et al, 2002).

Measures for Summative Testing

- *Effectiveness*: The precise criteria for successful achievement of each task should be used to determine the number of users successfully achieving each goal.
- *Efficiency*: The time taken to achieve the goal(s) should be measured from the time the user is asked to start to the time the goal has been achieved. The expected time to complete the task should be estimated from pretests, and the maximum time allowed to users before they are categorized as having failed should be at least three times the expected time.
- *Satisfaction*: questionnaire results.

The results of a summative test of the developed system can be documented using the Common Industry Format for Usability Test Reports (ISO, 2005), and compared with the usability requirements. (This format is too detailed for purely formative test reports, but NIST is coordinating a similar initiative to make recommendations on the formats that can be used for formative testing (IUSR, 2006).)

Usability Problems

The nature of any difficulties encountered by users can be recorded in order to identify usability problems that can be provided as feedback to design as in a formative evaluation.

Remote Usability Testing

In a remote usability test, the user works in their own environment, and their interaction may be observed remotely, either by video-conferencing and/or by use of a tool such as Morae that reproduces the user's screen remotely. Alternatively, users may just be asked to identify and report any problems encountered. Remote evaluation has been shown to produce reliable summative data (Petrie et al, 2006), but unless an audiovisual link is used, it is difficult to identify usability problems.

References

- Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, and B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 189-194). London, UK: Taylor and Francis.

- Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.) Human mental workload (pp.139-183). Amsterdam: North-Holland.
- ISO/IEC 25062 (2006). Software Engineering - Software product Quality Requirements and Evaluation (SQuaRE)-Common Industry Format (CIF) for Usability Test Reports.
- IUSR 2006. www.nist.gov/iusr (accessed July 2006)
- Kirakowski, J. 2000 Questionnaires in usability engineering. www.ucc.ie/hfrg/resources/qfaq1.html (accessed Jul 2006)
- Lewis, J.R. (2006) Sample sizes for usability tests: Mostly math, not magic. Interactions (in press).
- Medlock, M. C., Wixon D., Terrano, M., Romero R., Fulton B. (2002). "Using the RITE method to improve products: a definition and a case study." Usability Professionals Association, Orlando FL July 2002
- Perfetti, C., & Landesman, L. (2001). Eight is not enough. Retrieved July 4, 2006 from http://www.uie.com/articles/eight_is_not_enough/.
- Sauro, J. (2006) UI Problem Discovery Sample Size. http://www.measuringusability.com/samplesize/problem_discovery.php Retrieved July 2006.