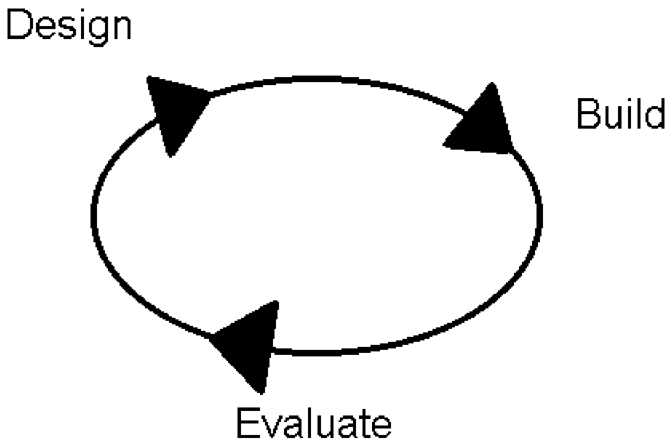
Evaluation and Usability Testing

A sound documentation development process should include rigorous evaluation of the documentation. This evaluation should be conducted at various stages as part of a cycle of designing, building, and evaluation. This is represented below.



Despite the great need for rigorously evaluating documentation, there is no doubt that far too little evaluation, especially empirical evaluation, is performed. The problem, of course, is limited resources, limited time in the schedule, limited amounts of salaried time the organization wishes to invest, lack of expertise in evaluation, and even lack of facilities and equipment. Savvy organizations, however, recognize the value of rigorous evaluation, build it into their schedules and budgets, and are richly repaid in terms of customer satisfaction and reduced technical support costs.

Evaluation can be divided into empirical and non-empirical evaluation. Empirical evaluation means that information is derived from actual users of the system or people who resemble users. Non-empirical evaluation means that the information comes from other sources, in particular expert opinion. First we will examination empirical evaluation.

Empirical evaluation

There are various forms of empirical evaluation. The most important are these:

- Questionnaires
- Interviews
- Focus groups
- Performance measurement
- Thinking-aloud protocols
- Field testing
- User logs

Often, it is best to perform more than one kind of empirical evaluation and to combine empirical and non-empirical evaluation. A major problem, however, is limited resources. This is discussed below.

Questionnaires

It takes considerable skill to design a useful questionnaire, to tabulate the responses efficiently, and to analyze the responses so that it yields meaningful results. On the other hand, a questionnaire is a means to collect data from many people without spending large amounts of time or money.

You can create questionnaires with numerical and yes/no questions (closed-ended questions) or with open-ended questions, where uses formulate their own responses. Closed-ended questions can be analyzed with relatively little effort, but the results may be less meaningful than open responses. One problem with open questions is that users are less likely to respond to the questionnaire because of the extra time required. A sound plan is to rely mainly on closed questions and conclude with a few open questions. Web-based questionnaires are easy to set up, are convenient for users, provide anonymity if desired, and do at least some of the tabulation automatically.

A major limitation of all questionnaires and with many other kinds of empirical evaluation is that you are just asking people's for their opinion. You are not observing or measuring actual performance. Users are often poor judges of documentation and of their own performance.

Interviews

In contrast to questionnaires, interviews are conducted face-to-face or by telephone. This results in back-and-forth interaction between the user of the documentation and the interviewer. Interviews can be open- or closed-ended, but if one is going to go to the trouble of actually interviewing users, it makes sense get the richer data provided by open questions. One problem is how to capture the user responses. Tape recording the users is very handy, but it can inhibit users.

Focus groups

Focus groups are discussion groups in which a moderator solicits responses from a group of users. The idea behind a focus group is that more and better information is generated when users listen to and respond to one another. One person's comment is apt to elicit related comments from other users, and it may become apparent that many of the participants are having the same problem. It is also possible, of course, that users will disagree, and it is possible that certain people in the group will exercise undue influence over others. Much of the success of a focus group depends on the skill of the moderator. If funds permit, it is good to hire a professional focus-group moderator, as long as the moderator understands enough about the issues that will be discussed.

A focus group, a questionnaire, and an interview will clearly be most effective when users have used the system recently. Therefore, it makes a great deal of sense to employ these evaluation methods as a follow-up to evaluation methods that entail using the system.

Performance measurement

Performance measurement generates actual data from users as they work with a system and its documentation. Normally, the user is given tasks to complete, and the evaluator measures relevant parameters such as percentage of tasks or subtasks successfully completed, time required for each task or subtask, frequency and type of errors, and duration of pauses, indications of user frustration, and the ways in which the user seeks assistance. This kind of performance measurement is what we usually mean by "usability testing."

There are many ways to design a usability test, and it is crucial to devise an appropriate test methodology. First and foremost, you need to know what you are trying to find out. For example, are you trying to find out the overall effectiveness of a help system? Or, are you interesting in one design issue, such as size of headings or use of color? Unfocused "fishing expeditions" are usually unproductive. Some of the crucial issues in devising a test is choosing the participants, deciding how many participants you need, handling participants so that you don't bias your data, and deciding when to you may want to discount the data you get from particular participant—perhaps a person who is clearly in too great a hurry or who turned out not to fit your participant profile. Other crucial issue are the nature of the tasks and test materials and possible time limits. How realistic are the tasks you are asking the user to perform? How closely do the testing conditions resemble the conditions of actual use?

One very difficult problem in usability testing and other forms of evaluation is the relationship between the documentation and the software (or other system). Should you ask users to read the documentation or to work in their normal way, consulting the documentation only if they wish to? If users choose not to consult the documentation, you won't learn about it. If you ask users to read the documentation, you are creating an artificial situation.

Thinking-aloud protocols

A thinking aloud-protocol is a kind of usability test in which the user is asked to continuously explain what he or she is thinking. The benefit here is that you can more readily understand the user's mental processes and, especially, what is causing a problem. After all, knowing that a particular problem exists of not very helpful if you can’t find out why the problem exists and get some ideas for addressing it.

One difficulty with thinking-aloud protocols is keeping the user talking. This can usually be achieved with occasional prompts. A bigger problem is studying the enormous amounts of data that is collected. This takes a great deal of time and patience. You may need to review many hours of video recording or perhaps have the participants’ commentary transcribed for easier examination—but this is expensive. Another problem is the artificiality of this kind of evaluation. Some usability experts question whether need to keep talking alters the way the participant thinks and works. Certainly, you can’t collect data about task completion times when the participants are being slowed down by having to articulate all their thoughts.

One variation on thinking-aloud protocols is to videotape participants working with the system and its documentation and then to show them the videotape (as soon as possible) and to ask them to comment on everything they were thinking as they were working. Participants can usually recall what they were thinking, and they are not distracted by the need to keep talking while they are performing the task. Furthermore, the evaluator can decide what parts of the test session were most important and can ask the participants to comment only on those parts. This may result in more cooperative participants. As you can imagine, there are significant benefits in videotaping all kinds of usability tests and other empirical evaluations.

Field testing

Field testing refers to observing people in their actual work environment. Normally, you observe them doing their actual work. You get to see what they really do. Because field testing is highly realistic, your findings are apt to be more meaningful than data collected when people perform artificial tasks in a lab setting.

Field testing, however, poses some real challenges. Apart from getting permission to watch people as they work, you may have to watch them for a long time before the work they do sheds light on the issues you care most about. When you create tasks in a laboratory usability test, you can focus on the issues you care most about.

User logs

A user log is a technology for recording a user’s interaction with the system. Typically this means installing a special (“instrumented”) software application on the user’s machine that “watches” everything the user does and records each word the user types and each command the user chooses. Logging is easier for web-based applications

There are various ways to employ user logs, but they are especially valuable for conducting long-term empirical evaluations of user behavior. A major limitation of most usability tests is that they last—at the most—a few hours. How can we really study how a user gradually learns a complex software application in such a short time? But if a user agrees to user logging, it is possible to collect data covering a period of many days or weeks. Potentially, you can compare the logs of multiple users.

The drawback here is that the data can be hard to interpret. For example, you must infer, or try to infer, what the user was trying to do in order to understand what problems the user encountered. Much depends on the kinds of reports the user log can create. Does the software record time information or just the succession of actions? Can the software tally and categorize the data into meaningful categories? Even with these drawbacks, user logs can be extremely valuable because they allow evaluators to get data that is not available in any other ways.

Non-empirical evaluation

Non-empirical evaluation consists of advice and other information that does not come from users or potential users of your system and its documentation. For this reason, it is not “data.” Non-empirical information normally derives directly or indirectly from experts. Non-empirical information can certainly be valuable, and it’s almost always easier and cheaper to obtain than empirical data. Also, certain forms of empirical testing may not be possible in the early stages of a project. There are two main kinds of non-empirical information

- Expert opinions

- Published literature including industry guidelines

Expert opinion

A documentation expert can often provide a very valuable critique or answer specific design questions. Whereas usability tests often focus on specific issues, an expert can holistically evaluate a help system or manual and pinpoint specific problem areas. Even if the expert charges significant consulting fees, the cost is likely to be less than empirical testing. Furthermore, an expert evaluation can be performed quickly.

One problem with this form of evaluation is deciding who is really an expert and who has expertise on the kind of documentation you are creating. An expert who has focused on the documentation for consumer applications may have little to say about documentation for developers. More important, there is often a significant gap between what an expert notices and finds fault with and what actually causes problems for users. Ultimately, even the most qualified expert is only guessing about how users will respond to a system and its documentation.

Published literature including industry guidelines

Instead of bringing in a flesh-and-blood expert, you can make use of expertise that has been published. This approach is cheaper, though probably not faster, than hiring a consultant. There is a lot of valuable information in the published literature, much of it derived from sound empirical studies. The most convenient form of published literature consists of guidelines (often called heuristics) and more elaborate guidelines called design pattern.

The problems with using literature sources are that the appropriate sources must be located, read, and applied to your project. This means that the person conducting the evaluation must be both energetic and knowledgeable. Perhaps the biggest pitfall lies in applying the literature to your project. It is very easy to use the published literature to convince yourself that your documentation is first rate.