Usability Metrics

nngroup.com/articles/usability-metrics/

World Leaders in Research-Based User Experience

This article is just exploring some usability metrics you can apply. Please research about the other metrics, and find out what else you can include?
Can you include number of clicks to complete a task for example?

Usability can be measured, but it rarely is. The reason? Metrics are expensive and are a poor use of typically scarce usability resources.

Most companies still under-invest in usability. With a small budget, you're far better off passing on quantitative measures and reaching for the low-hanging fruit of qualitative methods, which provide a much better return on investment. Generally, to improve a design, insight is better than numbers.

However, the tide might be turning on usability funding. I've recently worked on several projects to establish formal usability metrics in different companies. As organizations increase their usability investments, collecting actual measurements is a natural next step and does provide benefits. In general, usability metrics let you:

- **Track progress between releases.** You cannot fine-tune your methodology unless you know how well you're doing.
- **Assess your competitive position.** Are you better or worse than other companies? *Where* are you better or worse?
- **Make a Stop/Go decision before launch.** Is the design good enough to release to an unsuspecting world?

- **Create bonus plans for design managers and higher-level executives.** For example, you can determine bonus amounts for development project leaders based on how many customer-support calls or emails their products generated during the year.

How to Measure

It is easy to specify usability metrics, but hard to collect them. Typically, usability is measured relative to users' performance on a given set of test tasks. The most basic measures are based on the definition of usability as a quality metric:

- success rate (whether users can perform the task at all),
- the time a task requires,
- the error rate, and
- users' subjective satisfaction.

It is also possible to collect more specific metrics, such as the percentage of time that users follow an optimal navigation path or the number of times they need to backtrack.

You can collect usability metrics for both novice users and experienced users. Few websites have truly expert users, since people rarely spend enough time on any given site to learn it in great detail. Given this, most websites benefit most from studying novice users. Exceptions are sites like Yahoo and Amazon, which have highly committed and loyal users and can benefit from studying expert users.

Intranets, extranets, and weblications are similar to traditional software design and will hopefully have skilled users; studying experienced users is thus more important than working with the novice users who typically dominate public websites.

With qualitative user testing, it is enough to test 3–5 users. After the fifth user tests, you have all the insight you are likely to get and your best bet is to go back to the drawing board and improve the design so that you can test it again. Testing more than five users wastes resources, reducing the number of design iterations and compromising the final design quality.

Unfortunately, when you're collecting usability metrics, you must test with more than five users. In order to get a reasonably tight confidence interval on the results, I usually recommend testing 20 users for each design. Thus, conducting quantitative usability studies is approximately four times as expensive as conducting qualitative ones. Considering that you can learn more from the simpler studies, I usually recommend against metrics unless the project is very well funded.

Comparing Two Designs

To illustrate quantitative results, we can look at those recently posted by Macromedia from its usability study of a Flash site, aimed at showing that Flash is not *necessarily* bad. Basically, Macromedia took a design, redesigned it according to a set of usability guidelines, and tested both versions with a group of users. Here are the results:

	Original Design	**Redesign**
Task 1	12 sec.	6 sec.
Task 2	75 sec.	15 sec.
Task 3	9 sec.	8 sec.
Task 4	140 sec.	40 sec.
Satisfaction score*	44.75	74.50

*Measured on a scale ranging from 12 (unsatisfactory on all counts) to 84 (excellent on all counts).

It is very rare for usability studies to employ tasks that are so simple that users can perform them in a few seconds. Usually, it is better to have the users perform more goal-directed tasks that will take several minutes. In a project I'm working on now, the tasks often take more than half an hour (admittedly, it's a site that needs *much* improvement).

Given that the redesign scored better than the original design on all five measures, there is no doubt that the new design is better than the old one. The only sensible move is to go with the new design and launch it as quickly as possible. However, in many cases, results will not be so clear cut. In those cases, it's important to look in more detail at *how much* the design has improved.

Measuring Success

There are two ways of looking at the time-to-task measures in our example case:

- Adding the time for all four tasks produces a single number that indicates "how long it takes users to do stuff" with each design. You can then easily compute the improvement. With the original design, the set of tasks took 236 seconds. With the new design, the set of tasks took 69 seconds. The improvement is thus **242%**. This approach is reasonable if site visitors typically perform all four tasks in sequence. In other words, when the test tasks are really subtasks of a single, bigger task that is the unit of interest to users.

- Even though it is simpler to add up the task times, doing so can be misleading if the tasks are not performed equally often. If, for example, users commonly perform Task 3 but rarely perform the other tasks, the new design would be only slightly better than the old one; task throughput would be nowhere near 242% higher. When tasks are unevenly performed, you should compute the improvement separately for each of the tasks:
    - Task 1: relative score 200% (improvement of 100%).
    - Task 2: relative score 500% (improvement of 400%).
    - Task 3: relative score 113% (improvement of 13%).
    - Task 4: relative score 350% (improvement of 250%).

  You can then take the geometric mean of these four scores, which leads to an overall improvement in task time of **150%**.

Why do I recommend using the **geometric mean** rather than the more common arithmetic mean? Two reasons: First, you don't want a single big number to skew the result. Second, the geometric mean accounts fairly for cases in which some of the metrics are negative (i.e., the second design scores less than 100% of the first design).

Consider a simple example containing two metrics: one in which the new design doubles usability and one in which the new design has half the usability of the old. If you take the arithmetic average of the two scores (200% and 50%), you would conclude that the new design scored 125%. In other words, the new design would be 25% better than the old design. Obviously, this is not a reasonable conclusion.

The geometric mean provides a better answer. In general, the geometric mean of *N* numbers is the *N* 'th root of the product of the numbers. In our sample case, you would multiply 2.0 with 0.5, take the square root, and arrive at 1.0 (or 100%), indicating that the new design has the same usability as the baseline.

Although it is possible to assign different weights to the different tasks when computing the geometric mean, absent any knowledge as to the relative frequency or importance of the tasks, I've assumed equal weights here.

Summarizing Results

Once you've gathered the metrics, you can use the numbers to formulate an overall conclusion about your design's usability. However, you should first examine the relative importance of performance versus satisfaction. In the Macromedia example, users' subjective satisfaction with the new design was 66% higher than the old design. For a business-oriented website or a website that is intended for frequent use (say, stock quotes), performance might be weighted higher than preference. For an entertainment site or a site that will only be used once, preference may get the higher weight. Before making a general conclusion, I would also prefer to have error rates and perhaps a few additional usability

attributes, but, all else being equal, I typically give the same weight to all the usability metrics. Thus, in the Macromedia example, the geometric mean averages the set of scores as: `sqrt` (2.50*1.66)=2.04. In other words, the new design scores 204% compared with the baseline score of 100% for the control condition (the old design).

The new design thus has 104% higher usability than the old one.

This result does not surprise me: It is common for usability to double as a result of a redesign. In fact, whenever you redesign a website that was created without a systematic usability process, you can often improve measured usability even more. However, the first numbers you should focus on are those in your budget. Only when those figures are sufficiently large should you make metrics a part of your usability improvement strategy.

More Case Studies

See full report on the return-on-investment (ROI) from usability for many more examples of before–after usability metrics. For even more depth, see the full-day course on Measuring User Experience,