Beyond Time and Errors – Position Statement

George Robertson
Microsoft Research
One Microsoft Way
Redmond, WA 98102
425-703-1527
ggr@microsoft.com

ABSTRACT

In this position statement, I describe my background in evaluation of visualizations along with a characterization of current techniques and their problems.

Keywords

Information Visualization, Usability Evaluation, Utility Evaluation, User Studies.

1. INTRODUCTION

My introduction to evaluation of user interfaces with user studies was in 1975 at Xerox PARC. I learned the basic techniques of measuring time to task, errors, subjective preference, video logging, keystroke models, and system instrumentation. In 1976, I used those same techniques to start a user study lab at Carnegie-Mellon University in the Computer Science Department.

In 1988, when I joined Xerox PARC and began working on the first Information Visualization system, we talked about using those same techniques and decided against it. We believed that time and error studies would not be sufficient and may, in fact, be misleading. We needed some broader technique that would take into account the whole sensemaking system. We did not succeed in developing a new evaluation methodology. We used keystroke model analysis and memory bandwidth models as a substitute for the traditional studies.

In 1996, I joined Microsoft Research to continue work on Information Visualization research. I returned to the use of time, error, and subjective preference studies as part of an iterative design process. Almost all of the technology we have developed in the last 12 years has been studied using these techniques. However, I have been aware for many years that these studies fail to characterize or evaluate the utility of the visualizations.

2. USABILITY EVALUATION

I have used the following techniques for usability evaluation of both user interfaces and visualization systems. In every one of these techniques, the choice of the right tasks for the study is critical. The tasks must be ecologically valid, and this is not always easy to determine. Even harder, the tasks must not be picked with some bias toward what the system being tested is capable of or excels in. The wrong choice of tasks will potentially mislead the designer.

In addition, the right choice of participants is critical. The participants must be representative of the real end-user. Choosing co-workers or students may be easy, but may not represent the target user, and hence will potentially mislead the designer.

2.1 Heuristic Evaluation – Keystroke Models

In some cases, effective evaluation can be performed simply by counting keystrokes, mouse operations, and mental operators needed to accomplish some task. This technique can work during early formative stages of design.

2.2 Discount Usability Testing

It has been repeatedly shown that a small number of participants (on the order of 5) is sufficient to discover many (if not most) of the basic usability issues with a new visualization. This technique can also work during early formative stages of design, even using paper prototypes.

2.3 Lab Studies

The traditional user study measures time to task, errors, and subjective user satisfaction. These studies typically require 12 to 20 participants. They can be used to test specific features (i.e., test performance with and without a particular feature), or can be used to compare a new visualization against an existing system.

2.4 RITE Studies

One problem with the traditional lab study is that the system being evaluated is held constant throughout the study, even if obvious problems become apparent after the first few participants. A new technique, called Rapid Iterative Technology Evaluation (RITE), is a variant of traditional lab studies that uses discount usability testing principals. After the first 2-3 participants, a list of usability issues is evaluated by the design team. If there is agreement on the nature of a problem, its cause, and a fix for it, the system is changed. The second round of 2-3 participants then evaluates the modified system. This technique allows the remaining participants to evaluate fixes as they are applied.

2.5 Field Studies

Another technique that is often used is to study a group of participants before and after introducing a visualization. The visualization system is often instrumented to gather data on its use. These studies are often time consuming, but yield deeper insight into the usability and in some cases the utility of a system.

3. UTILITY EVALUATION

While usability evaluation is important and can lead to visualizations that are more usable, they can miss a very fundamental issue. Any effective visualization must be helping users solve some real-world problems. That is, they must be useful as well as usable. What is missing is a proper characterization of the real-world problems being solved and an evaluation technique against those problems.

3.1 Insight-based Evaluation

Richard Hamming is quoted as saying “the purpose of computing is insight, not numbers.” This leads to a generic evaluation of visualization techniques by counting the number of insights participants get during some fixed period of time using a visualization. While this is a promising technique, it suffers from several potential problems. First, it requires domain experts as participants. Second, some insights are more important than others, so a subjective evaluation must be done to determine the value of the insights. Third, the ground truth is rarely known – that is, no one really knows what the correct insights are.

3.2 Task Identification & Heuristic Utility Evaluation

I believe that the fundamental problem is that we have not carefully characterized the problems we are setting out to solve. We sometimes have solutions in search of a problem. I believe the first step in any visualization design should be to spend time with the end user and understand exactly what problems they are trying to solve. Understanding the tools they currently use, and how effective or ineffective those tools are, is a legitimate way to guide the design process. To design a useful visualization requires a careful characterization of the user tasks. In turn, that characterization of user tasks can drive the choice of tasks for usability studies. It can also drive heuristic utility evaluation. I would argue that counting insights is not sufficient if those insights are not used in solving the problem at hand. In addition, understanding how the visualization is used in the sensemaking cycle can lead to greater insight into its usefulness. Some combination of task characterization, insight-based evaluation, sensemaking analysis, and heuristic utility evaluation is needed. That is the key problem that needs to be solved.