USABILITY TESTING OF SOFTWARE FOR ASSESSING COMPUTER USAGE SKILLS

Heidi Horstmann Koester
Koester Performance Research
Holland, MI

William W. McMillan
Eastern Michigan University
Ypsilanti, MI

Background

We have developed an initial prototype of software for measuring the keyboarding and mouse usage skills of people with disabilities. This software, called Compass, allows an evaluator to assess an individual's computer input skills and compare performance across different devices and time periods. The main features of this prototype have been described elsewhere (Koester and McMillan, 1997).

Research Goal

Usability testing has been employed throughout the project to identify user needs and evaluate how well our software meets those needs. The primary methods of gathering and incorporating user feedback into the development of the prototype are described below.

Defining Measurable Objectives

A key element of usability testing is to define measurable behavioral objectives that provide concrete usability benchmarks (Gould, 1988). To help formulate measurable objectives, a survey was developed which included questions on the respondent's background; goals and time spent in a typical client evaluation; interest in a quantitative assessment tool; rated importance of 23 different features of such a tool; and acceptable learning time, usage time, and cost. The survey was placed on a web site, and 37 computer access clinicians completed it.

Measurable objectives for learning time and evaluation time were defined from the survey re-sponses. For learning time, 40% of responders were willing to spend 1-2 hours to learn the software. 35% deemed 31-60 minutes acceptable, with the remaining responses at 30 minutes or less. A reasonable target, then, was defined as a learning time of 31-60 minutes.

For evaluation time, the average responder said that they spent 31-60 minutes evaluating keyboarding and pointing skills under current practice. The majority (57%) were willing to spend a little longer on evaluations using the Compass software, but the average response rounded down to "not longer", which suggests that a better target is that evaluations with the Compass system should not take longer than current practice. Our target for evaluation time, then, was defined at 31-60 minutes.

Usability Test #1

A usability test was conducted on an initial prototype of the Compass system, which incorporated a fairly complete evaluator interface, as well as early implementations of two keyboarding tests. The goals were to determine whether a typical user could meet the defined targets for learning and evaluation time, measure overall satisfaction with the software, and observe errors and misunderstandings.

Methods. Eight experienced clinicians participated. Each performed six pre-defined tasks with the system, which were printed out on paper for reference during the test. Short, dynamic help messages were available at the bottom of each screen, but otherwise there was no on-line help implemented in the system nor were there written instructional materials. Subjects were asked to try to solve any problems on their own, with the option of asking questions of the experimenter if they really felt stuck.

Each subject action and its associated time was recorded. Data were analyzed to determine for each task: the successful completion rate, the completion time, and the number and type of errors made. Any comments were also recorded for additional insight into possible problems. Subjects completed a post-test survey which assessed their level of agreement with 12 different statements about the prototype. Answers were on a 1 - 5 scale, from "strongly disagree" to "strongly agree." A score significantly greater than 3.6 is considered significant for positive survey questions, and a score significantly lower than 2.4 is considered significant for negative survey questions (Nielsen, 1995).

Before testing began, the specific objectives defined from the web survey were converted to measurable targets for the defined tasks in this test. For learning time, the target was independent use after a guided experience lasting 31 - 60 minutes. The six tasks were designed to take approximately one hour, with little or no help from the experimenter. A positive response to the post-test question of whether subjects "felt capable of using Compass with a client after this experience" was one indicator of successful achievement of the learning time goal. Other questions, asking whether Compass was "very easy to learn" and whether its use was "very

frustrating", were additional indicators of perceived learning ease. A final set of indicators was that subjects should be able to complete Tasks 1 and 2 with no experimenter help, providing confidence that the basics of system use can be achieved quickly, and an overall completion rate of 100%, with no more than two interventions per subject by the experimenter.

For evaluation time, the target for assessment of keyboarding and pointing skills with the Compass system was 31-60 minutes. Since in this initial prototype, only keyboarding tests were implemented, the target was divided in half, to 16-30 minutes. Tasks 4-6 were designed to be a reasonable representation of an evaluation, requiring three runs of the Sentence test under three different configurations and interpretation of the resulting reports. Therefore, the primary criterion for achieving the evaluation time target was an average time under 30 minutes across Tasks 4-6. A secondary criterion was that Tasks 1 and 2, which represent single tests administered to a client, should each take less than 7 minutes to complete (allowing 4 tests to be administered within the target time). Two survey questions were additional indicators of satisfactory evaluation time ("Compass would take longer than my current assessment methods" and "It is worth the effort to use Compass").

Results. For learning time, five of the six measurable objectives were met. Subject responses on the questions of independent use, ease of learning, and frustration easily exceeded target levels. All subjects were able to complete the tasks successfully, with only two instances across all subjects in which the experimenter gave a small hint. Both of these hints were to one subject on Task #2, so the objective of completing Tasks 1 and 2 without help from the experimenter was not fully met.

For evaluation time, five of the six measurable objectives were met. Subject responses on the question of whether use of Compass was worth the effort averaged 4.4, significantly greater than the target of 3.6. Subjects were less sure if use of Compass would take longer than current practice, with an average response of 2.0, which was not significantly different than 2.4. All measured time criteria were met. The average time for Tasks 4 - 6 was 14.8 minutes, which was significantly below the target level.

Subjects committed an average of 7.6 errors across the six tasks. In all but two instances, subjects were able to recover easily from their errors with no experimenter help. These errors were traced to 22 distinct usability problems. Nine of the problems occurred with a frequency greater than 50%. Most of these problems were related to Compass' ability to let the evaluator define and run a group of tasks with a client (in one or more sessions), as well as change the list of tasks or their configurations at any time. This feature was well-liked by evaluators, and is a key to Compass' power and efficiency in real-world use, but it does require the evaluator to manage a list of tasks and their configurations.

On the basis of these results as well as subjects' comments, the Compass interface was revised, to reduce or eliminate as many usability problems as possible. While these problems did not in general prevent subjects from reaching the behavioral objectives, they did represent opportunities for improving the interface.

Usability Test #2

Methods. A second usability test was conducted once the revisions to the Compass interface were complete and two new pointing tests were implemented. Ten participants were solicited from clinicians in the U.S. and Canada who had expressed interest in the Compass project. Participants evaluated the software by performing suggested and self-defined tasks and completed the post-test survey as well as some open-ended questions.

Data collection focused on survey question responses and qualitative feedback provided by the evaluators. Since participants were geographically scattered, it was not possible to observe the time required for learning and evaluation. Therefore, measurable objectives for user satisfaction were defined as an average score significantly greater than 3.6 for "positive" survey questions and an average score significantly lower than 2.4 for "negative" survey questions. The qualitative feedback was carefully analyzed and collated to identify usability problems and other suggestions for enhancements to the system.

Results. Subjects in both usability tests completed the same post-test survey, and statistical analysis showed that responses were not significantly different for the two subject groups. Therefore, responses were collapsed across all subjects to gain the benefit of a larger subject pool.

Table 1 shows the average response to each survey question. Responses met the target level for 8 of the 12 questions, which suggests a relatively high level of user satisfaction overall. Responses to the other 4 questions indicate the following: subjects did not agree on whether use of Compass would take longer than current practice;

the response time of the system could be improved; the client tasks could be made more motivating through the addition of color, animation, and other features; and planned additional keyboarding tasks should be implemented.

Discussion

These usability tests have provided invaluable information as we develop the Compass software. Quantifying even some user needs and verifying that the system meets those needs gives increased confidence that clinicians may ultimately find Compass to be a useful tool. Additionally, while beyond the scope of this paper, perhaps the richest source of information was the qualitative feedback provided by participants. Our next development cycle will focus on incorporating these clinician comments into the system.

References

Koester, H., McMillan, W. (1997). Software for Assessing Computer Usage Skills. *Proc. of RESNA '97*, 354-356.

Gould, J. (1988). How to design usable systems. In *Handbook of Human-computer Interaction.* Elsevier Science Publishers.

Nielsen, J. (1995). *Usability Engineering*. Boston: AP Professional.

Acknowledgments

This work was funded by the National Institutes of Health, grant #1R41-NS36252-01, as a Phase I STTR award to Koester Performance Research. Many thanks to the participating clinicians for their time and thoughtful insights.

Heidi Horstmann Koester
368 Oak Harbor Ct.
Holland MI 49424
hhk@umich.edu

Survey Question	Ave. Response	95% C.I.	Met Goal?
It was very easy to learn how to use Compass.	4.1	[3.9, 4.4]	√
Using Compass was a very frustrating experience.	1.6	[1.2, 1.9]	√
I feel I am capable of independently using Compass with a client after this experience.	4.3	[3.7, 4.8]	√
The reports of the results were clear.	4.2	[3.7, 4.6]	√
Compass is very pleasant to work with.	4.2	[3.9, 4.5]	√
Compass would probably take longer to use than my current assessment methods.	2.3	[1.7, 2.8]	No
I am likely to use Compass routinely for client assessments.	4.2	[3.8, 4.6]	√
It is worth the effort to use Compass.	4.4	[4.1, 4.7]	√
The measures Compass provides are accurate indicators of a client's keyboarding skill.	3.8	[3.4, 4.2]	No
I understood how to do the client tasks (Single Letter, Sentence, Aim, and Menus)	4.3	[3.7, 4.9]	√
My clients would find the Compass tasks motivating.	3.3	[2.9, 3.8]	No
Compass seems to respond slowly.	2.8	[2.1, 3.5]	No

Table 1. Responses to post-test survey questions across the 16 subjects in Usability Tests #1 and #2.