FURTHER VALIDATION OF THE "AIM" TEST

FOR ASSESSING A USER'S MOUSE SKILLS

Heidi Koester, Edmund LoPresti, Richard Simpson
Koester Performance Research, University of Pittsburgh

INTRODUCTION

Compass software for computer access assessment includes eight skill tests that measure a user's performance for mouse use, text entry, and switch use. One of the most commonly used Compass tests is the Aim test, which examines target acquisition skill (i.e., the ability to click on an object). This study examined the test-retest reliability and construct validity of the Aim test, with 16 individuals who have physical impairments. Results provide confidence in the Aim test as a valid assessment tool.

BACKGROUND

The purpose of Compass software is to provide clear evidence about a user's ability to use various computer access options, such as input devices and display settings. This evidence helps an individual or practitioner determine which access solutions will best meet a user's specific needs. Compass has undergone extensive usability testing, demonstrating its ease-of-use [1], as well the accuracy of its timing and accuracy measurements [2].

A full understanding of its psychometric properties is also important, to ensure that Compass effectively fulfills its purpose. A prior study measured test-retest reliability, intra-test reliability, internal consistency, and construct validity for six Compass tests, with excellent results [3]. Recently, our research on a related project yielded data that allowed us to re-examine test-retest reliability and construct validity for the Aim test.

Test-retest reliability represents how consistent test results are when the same test is administered multiple times in a row. While successive results are unlikely to be identical, a wide variation suggests that the test may not be a very reliable measuring tool. Reliability is particularly important for Compass since a common application is to administer a test multiple times, trying a different input device each time, to determine the device that provides the best performance for the user. If the test itself is unreliable, it is impossible to draw conclusions about the effect of input device.

Construct validity reflects how well the test measures what is intended to measure. With high construct validity, results on Compass tests would generalize well to the user's performance on real world computer tasks. Conversely, low construct validity would severely limit the usefulness of the Compass test results.

Hypothesis

The test-retest reliability and construct validity for the Compass Aim test will be high enough to make it a valid instrument for assessing an individual's skill with a pointing device.

METHODS

Overview

The Compass Aim test presents a series of single targets on the screen, which the user selects by clicking on each target in turn. Individuals followed a protocol where they performed the Aim test twice in a row. The protocol also included several 'real-world' Windows target acquisition tasks, such as clicking on a scrollbar button and selecting a menu item. These activities were part of a larger protocol for a related research study; in this paper we analyze the data that is relevant to the psychometric properties described above.

Participants

Across two similar studies, 16 unique individuals participated. All participants had some prior computer experience, could see and interpret the test stimuli, and had a physical impairment that affected their ability to use a pointing device. Pointing devices were assigned to participants to match their own input devices, as follows: mouse (N=7), trackball (N=6), joystick (N=1), head mouse (N=1), MouseKeys (N=1). Clinical diagnoses included cerebral palsy (N=6), cervical spinal cord injury (N=4), brain injury (N=4), multiple sclerosis (N=1), and muscular dystrophy (N=1).

Protocol

Participants completed a questionnaire regarding basic demographic information, the nature of their disability, computer experiences, and input devices that they currently use. They performed the Compass Aim test twice. Each test presented 32 targets. Half of the targets were 16 pixels square, and half were 32 pixels square, presented in random order and at random locations. Compass measured the time required to select each target.

In the second part of the protocol, participants followed a written script to perform four 'real-world' target acquisition tasks in Windows. The tasks were: press the scrollbar button to move to a particular location in a document, minimize a window, maximize a window, and select the File/Exit menu item. The time required to complete each task was manually measured from reviewing the video recordings of the user's computer screen.

Data Analysis

To measure test-retest reliability for the Aim test, the intra-class coefficient (ICC) was calculated using the timing measurements from the first and second repetitions of the test. ICCs between 0.80 and 1.00 are considered to represent high test-retest reliability; those between 0.60 and 0.79 are "moderately reliable." [4]

The mean percent difference in selection times for the first and second tests was also calculated, with a difference of less than 15% as our target for high reliability. The signed percent difference was calculated as (Time2 – Time1)/Time1 * 100. The absolute percent difference is the absolute value of the signed percent difference, and measures the amount of deviation, whether positive or negative, of the retest relative to the initial test.

For construct validity, subjects' average selection times for the first Aim test were compared to the average selection times across the four 'real-world' tasks. We calculated the correlation between these two variables as an indicator of how closely the Aim results reflect 'real-world' performance.

RESULTS

Test-Retest Reliability

Figure 1 shows how the average selection times in the Aim retest compared to the initial test. 13 individuals completed both tests. The ICC was 0.963, which is significant at $p < 0.001$ and has a 95% lower limit of 0.885.

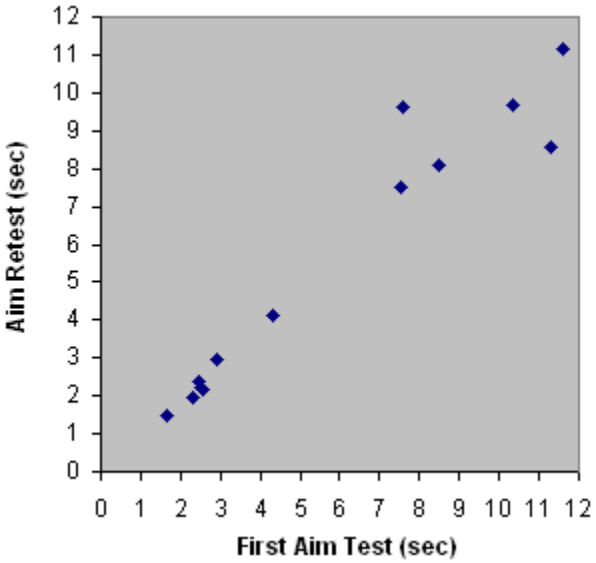

Figure 1. Selection Times in Aim Retest vs. First Aim Test.

Raw retest times were slightly faster than the initial test, with an average signed percent difference of -5.8%. The average deviation (absolute percent difference) was 10.1%.

Construct Validity

Figure 2 shows how the average real-world task times compared to the selection times in the first Aim test. All 16 participants are

represented. The correlation between real-world and Aim times was 0.912, significant at $p < 0.001$. Squaring this correlation, we see that 83% of the variance in real-world task times is explained by the Aim test scores.

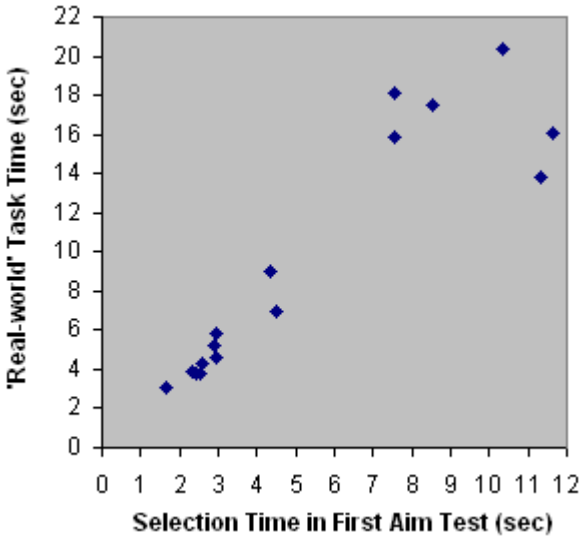

Figure 2. 'Real-world' Target Acquisition Time vs. Selection Time in the First Aim Test.

DISCUSSION

The results for both test-retest reliability and construct validity strongly support our hypothesis. The high ICC and low percent difference suggest that the Aim test has excellent reliability. The strong relationship between Aim scores and real-world task performance further suggests a high degree of construct validity.

These results are consistent with those from our earlier psychometric study [3]. In that study, the ICC for the Aim test was 0.976 (vs. 0.958 here) and the absolute deviation was 14.0% (vs. 10.1% here). This replication reinforces the strength of these results.

One interesting aspect to these results is the large range in performance across these 16 individuals. An experienced mouse user without physical impairment can typically select an Aim target in about 1 second. Times for these users ranged from just under 2 seconds up to 12 seconds.

Additionally, there are no Aim test times between 5 and 7 seconds in this data set. The scatter in the graphs appears to have more variation for Aim scores above 7 seconds. While more data would be required to draw a clear conclusion, this does make intuitive sense; times above 7 seconds represent significant difficulty, and performance in that range may be more variable.

While these results are encouraging, it should be noted that this study only examined the Aim test, rather than all eight Compass tests. Our previous study, which included six Compass tests, showed similar results across all the tests [3], but it would be desirable to replicate those results with more than just the Aim test. Additionally, while we referred to the scripted Windows tasks as 'real-world' tasks, they may not have shared all of the characteristics of true day-to-day actions. Participants did not have to think about what action they wanted to take, since that was written out for them. But with respect to the target acquisition aspect of the task (which is the aspect of most interest here), the scripted tasks were exactly like real Windows tasks.

A common and important concern of Compass users is how closely the results mirror performance in real, day-to-day computer tasks. These results, combined with those from our earlier psychometric study, provide strong confidence that Compass is a reliable and valid assessment tool for computer access.

ACKNOWLEDGEMENTS

Many thanks to the participants for their time and effort. This work was funded by a Phase II SBIR Award, National Institutes of Health.

REFERENCES

[1] Ashlock, G., Koester, H., LoPresti, E., McMillan, W., and Simpson, R. "User-centered design of software for assessing computer usage skills", in Proceedings of RESNA 2003 Conference, Arlington, VA: RESNA Press, 2003.

[2] Koester, H., LoPresti, E., Simpson, R. "Measurements validity for Compass assessment software", in Proceedings of RESNA 2006 Conference, Arlington, VA: RESNA Press, 2006.

[3] Koester, H., Simpson, R., Spaeth, D., and LoPresti, E. "Reliability and validity of Compass software for access assessment", in Proceedings of RESNA 2007 Conference, Arlington, VA: RESNA Press, 2007.

[4] Mazer, B., Dumont, C., and Vincent, C. "Validation of the assessment of computer task performance for children", Technology and Disability, 15: 35-43, 2003.