Reliability and Validity of Compass Software for Access Assessment

Heidi Horstmann Koester, PhD, Richard C. Simpson, PhD, Don Spaeth, PhD, Edmund F. LoPresti, PhD
Koester Performance Research; Human Engineering Research Labs, VA Pittsburgh Healthcare System

ABSTRACT
Compass software includes a variety of skill tests to assess users' computer usage abilities. This study examined the psychometric properties of the Compass tests, including test-retest reliability, intra-test reliability, internal consistency, and construct validity. 15 experienced computer users with physical impairments participated. Overall, the results confirm that Compass can be validly used as a computer access assessment tool.

Keywords: computer access, user-computer interface, measurement, psychometric properties, assistive technology

BACKGROUND
Compass is a software tool for professionals who provide services in computer access and augmentative communication. The software measures user performance in skills needed for computer interaction, such as keyboard and mouse use, navigating through menus, and scanning. Appropriate use of Compass helps a clinician to:

- diagnose difficulties with an existing interface;
- evaluate and compare the expected performance of potential access systems;
- plan training interventions;
- track changes in an client's abilities over time; and
- measure the effectiveness of an intervention.

The current version of Compass includes eight skill tests in three input device domains. During a test run, the speed and accuracy of user actions are recorded. Compass reports summarize results for each test and provide trial-by-trial detail if desired.

Compass has undergone extensive usability testing, demonstrating its ease-of-use [1], as well as testing to confirm the accuracy of its timing and accuracy measurements [2]. In this study, we examined some key psychometric properties of the Compass skill tests, in order to help clinicians use Compass more effectively and to reveal any areas in which test revisions should be made. A similar study performed on another assessment instrument (the Assessment of Computer Task Performance) helped guide our choice of methods for this work [3].

Research Goals
In order to better understand the measurement properties of the Compass assessment skill tests, we pursued four specific objectives.

1. Measure Test-Retest Reliability for individual Compass skill tests. This represents how consistent user scores are on successive administrations of a skill test. This would help clinicians better interpret data from successive Compass tests, because they would understand how much variation to expect from sequential order effects.

2. Measure Intra-Test Reliability. This represents how consistent user performance is across trials within a given Compass skill test. This would help us better understand how many trials are necessary in each test to get a reliable picture of the user's performance.

3. Measure Internal Consistency. This tells us whether Compass scores for different users differ because those users have different computer input skills, or because the skill tests are confusing or problematic in some way.

4. Measure Construct Validity. Examining the factor structure underlying Compass skill tests will help us determine the extent to which Compass is truly measuring the constructs of interest, specifically the abilities to perform text entry, pointing device, and switch tasks on the computer.

METHODS

Overview

Subjects completed a questionnaire about the nature of their disability, basic demographic information, computer experience, and input devices that they currently use. They then performed Compass tasks as follows:

(1) a single run of each of three Compass skill tests: Aim, Drag, and Switch;
(2) a repeat run of Aim, Drag, and Switch;
(3) a single run of each of three Compass skill tests: Letter, Word, and Sentence;
(4) a repeat run of Letter, Word, and Sentence.

Participants

15 athletes from the 2006 National Veterans Wheelchair Games participated. All participants had some prior computer experience, could see and interpret the test stimuli, and could use the text entry and pointing devices that were available at the test site. Input devices were assigned to participants to match their own input devices, and participants used their own accessories as needed, such as typing splints. All used a standard keyboard; pointing devices used were standard mouse (N=11), trackball (N=3), and trackpad (N=1). All participants were male, with an average age of 58 years old. Clinical diagnoses included cervical spinal cord injury (N=8), multiple sclerosis (N=4), thoracic spinal cord injury (N=1), and traumatic brain injury (N=1).

Compass Tests

Each Compass test presents a series of trials measuring a particular skill. Summary scores for each test are computed for the average speed and accuracy of performance across all trials in the test. Speed and accuracy for each trial within a test are also recorded. The Aim and Drag tests are two of Compass' pointing device tests. The Aim test presents a series of single targets on the screen, which the user selects by clicking on each target in turn. Drag asks the user to drag a target to a destination. The Switch test is one of the single switch tests in Compass. It measures how quickly a user presses a switch (in this case, the mouse button) in response to a prompt. The Letter, Word, and Sentence tests are the three text entry tests; each presents a text string (a letter, word, or sentence) for the user to enter. While the presentation of each test is customizable, the standard settings for each test were used throughout this protocol. Two Compass tests, the Menu and Scan tests, were not included due to time constraints for data collection.

Data Analysis

To measure test-retest reliability for each Compass 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", and those under 0.60 suggest poor reliability. The mean percent difference in performance scores for the first and second tests was also calculated, with a difference of less than 15% suggesting high reliability.

To measure intra-test reliability, the trials for each test were divided into two halves, or Blocks. A mixed model ANOVA analysis was performed on the timing measures to examine the main effect of Block, with Subject as a random effect to control for subject differences. A significant Block effect indicates that subjects' performance time was significantly different in different halves of the skill tests. Graphs of average time for each Block were also examined to see the nature of any differences.

Cronbach's alpha was calculated as a measure of internal consistency, using the timing data for all 6 tests in the first test repetition (Aim, Drag, Switch, Letter, Word, and Sentence). An alpha value between 0.6 and 0.8 was considered desirable, indicating that the Compass tests tend to cohere toward an overall indicator of computer input skills, without being too redundant.

To measure construct validity, a factor analysis was performed on the timing data for all 6 tests in both test repetitions. Factor analysis addresses these questions: How many components (constructs) are needed to represent these variables? What do these components represent? Overall, Compass skill tests are designed to measure computer usage skill, so this single construct might emerge in the factor analysis. Additionally, some skill tests relate to pointing device use (Aim, Drag, and Switch), and some relate to text entry (Letter, Word, and Sentence), so we might expect two components to emerge in the factor analysis.

RESULTS AND DISCUSSION

Test-Retest Reliability

As shown in Table 1, all but two measures tested met the 0.80 criteria for high test-retest reliability. In the Letter test, the overall selection time and key press time were of only moderate reliability.

Test	**Measure**	**ICC**	**LL**
Switch	Time	.882	.686
	Press	.872	.662
	Release	.866	.649
Aim	Time	.976	.932
	Reaction Time	.935	.823
Drag	Time	.915	.742
Letter	Time	.655	.245
	Press	.658	.248
	Release	.823	.558
Word	Typing Speed	.913	.760
Sentence	Typing Speed	.834	.375

Table 1. Reliability of Compass Tests. ICC: intraclass correlation coefficient, LL: lower limit of the confidence interval of the ICC.

Test-Retest Differences

Test-retest differences measure how much the time performance for the Retest differed from the time performance for the first Test. The signed percent difference for each subject was calculated as (Time2 -

Time1)/Time1 * 100. The absolute percent difference is the absolute value of the signed percent error, and measures the amount of deviation, whether positive or negative, of the Retest relative to the Test. Table 2 shows the signed and absolute differences, averaged across subjects for each test.

	Signed Difference, %		Absolute Difference, %	
Test	**Mean**	**SD**	**Mean**	**SD**
Switch	-7.61	19.42	16.63	12.62
Aim	-1.01	18.42	14.01	11.40
Drag	-7.11	12.05	12.13	6.45
Letter	-11.47	29.62	25.12	18.50
Word	-1.93	15.35	11.85	9.44
Sentence	-13.27	13.39	16.27	9.19
All Tests	-7.07	18.04	16.00	11.17

Table 2. Test-Retest Performance Differences for Compass Tests.

The signed percent differences suggest the second test averages a bit faster than the first test. From the absolute averages, a "typical" deviation on retest is about 15%, for all tests except Letter, which has a higher deviation of 25%. This represents the expected "noise" of the participant-test combination. When we are looking for a meaningful change from one test to another (for example, due to a change in input device), a rule of thumb is that it should exceed this typical "noise" level. Note that these measurements reflect the scenario where the user has never performed the test before, then performs it twice in a row. Increased test familiarity may lead to more consistent results on subsequent retests, but additional data are needed to test this possibility.

Intra-Test Reliability

Table 3 shows that, for the Switch and Sentence tests only, subjects' times were significantly different in the first half of the test as compared to the second. In both cases, the second half of the test was faster than the first half. For the other tests, there were no significant differences between the test halves. This suggests that the number of trials in the test could be reduced without an important loss of information.

Test	**Trials per Block**	**p-value for Block effect**
Aim	12	.752
Drag	12	.082
Switch	5	.002*
Letter	5	.221
Word	5	.496
Sentence	2	.034*

Table 3. Statistical comparison of times between two halves (Blocks) of each Compass test.

Internal Consistency

For Times across all tests in the first test repetition, Cronbach's alpha = 0.768. This meets the goal of an alpha between 0.6 and 0.8. Note that Cronbach's alpha measures how well a set of variables measures a single underlying construct. The moderately high alpha observed here suggests that the Compass tests hang together to indicate overall computer usage skills, but, because each test measures a separate skill,

results across different tests would not be expected to be extremely consistent, which is why the Cronbach's alpha score isn't higher.

Construct Validity

In the factor analysis, two components emerged, explaining a total of 86% of the variance in the timing measurements. Table 4 below shows how strongly each variable related to each of the two components. Scores closer to +1 or -1 show stronger relationships, while scores closer to 0 reflect weaker relationships. For Component #1, all of the Aim, Drag, and Switch measures had high component scores, while all of the text entry measures had low component scores. This suggests that Component #1 relates to the construct of pointing device skill. Similarly, Component #2 is primarily made up of the text entry tests of Letter, Word, and Sentence, while the pointing-related tests have only a weak presence in Component #2.

	Component	
Test	**1**	**2**
Aim Test	.920	-.292
Aim Retest	.913	-.314
Drag Test	.889	-.072
Drag Retest	.922	-.069
Switch Test	.876	-.231
Switch Retest	.847	-.334
Letter Test	.382	-.768
Letter Retest	-.025	-.907
Word Test	-.214	.906
Word Retest	-.220	.948
Sentence Test	-.267	.894
Sentence Retest	-.269	.900

Table 4. Component Matrix for Compass Timing Measures. (All measures in the factor analysis were in units of seconds, except for Word and Sentence, which were in words per minute. This is why Word and Sentence have positive component scores for Component 2, while Letter has a negative Component 2 score.)

CONCLUSIONS

These results indicate that the Compass skill tests have an appropriate level of test-retest reliability and that they actually do measure the computer skills that they are intended to measure.

REFERENCES

1. Ashlock, G., Koester, H.H., LoPresti, E., McMillan, W., and Simpson, R. (2003). User-centered Design of Software for Assessing Computer Usage Skills. *Proceedings of RESNA 2003 Annual Conference*, Atlanta, GA. Arlington, VA: RESNA Press.

2. Koester, H.H., LoPresti, E.F., Simpson, R.C. (2006). Measurement Validity for Compass Assessment Software. *Proceedings of RESNA 2006 Annual Conference*, Atlanta, GA. Arlington, VA: RESNA Press.

3. Mazer, B., Dumont, C., and Vincent, C. (2003). Validation of the Assessment of Computer Task Performance for Children. *Technology and Disability 15*: 35-43.

ACKNOWLEDGEMENTS
Many thanks to the participants for their time and effort and to the University of Pittsburgh for coordinating and performing the data collection.

ADDRESS
Heidi Koester
2408 Antietam Dr.
Ann Arbor, MI 48105
Email: hhk@kpronline.com