Technical Report:
Reliability of the Highlands Ability Battery

Submitted to:

The Highlands Company
Larchmont, NY

Submitted by:

Andrew G. Neiner, Ph.D.
Industrial/Organizational Psychologist
Atlanta, GA

September 2013

Introduction

The study was conducted as part of The Highlands Company's on-going effort to ensure that the Highlands Ability Battery (HAB) continues to meet professional standards of excellence. Specifically, this study examined the reliability of each work sample using two different methods: internal consistency and test-retest.

Reliability and Speeded Tests

Reliability refers to the consistency of scores obtained by the same persons when they are reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions (Anastasi & Urbina, 1997). In psychometric terms, reliability is the proportion of true score variance to error variance. There is no single best method to assess reliability, each method has its advantages and disadvantages depending on the type of test being assessed.

Gulliksen (1950) distinguished between tests that measure only knowledge, called power tests, and those that also measure cognitive processing speed, called speeded tests. A speeded test is one in which individual differences in test scores depend entirely on speed of performance. Scnipke and Scrams ( ) define speededness as the extent to which time limits affect examinees' test performance. Pure speeded are comprised of items with uniformly low difficulty and no one is expected to attempt all the items in the time allotted - given enough time, all testtakers would be expected to answer all the items correctly. On the other hand, a pure power test has a time limit that permits everyone to attempt all the items, and the difficulty of the items are steeply graded. In creating the notion of speededness, Swineford (1956) established a way in which this classification could be determined. If all examinees (99%) reach 75% of the items and all of the items are reached by 80% of the examinees, then the test may be considered unspeeded.

In actual practice, the distinction between speeded and power tests is one of degree, most tests depending on both power and speed in varying proportions (Anastasi & Urbina, 1997). This is the case with the HAB work samples where both speed and knowledge influence test performance. Because of the imposed time limits, processing speed is incorporated into the ability construct. The degree of speediness varies among the work samples. This hybrid approach is by design.

Reliability Measures and Test Speededness

There are two major types of reliability measures: test-retest and internal consistency. The former assesses the consistency of test scores over time while the latter assess the inter-item consistency. Internal consistency describes the extent to which all the items in a test measure the same concept or construct and hence it is connected to the inter-relatedness of the items within the test.

Test speededness needs to be considered when choosing an appropriate reliability measure. Because examinees without enough time will often either hurry through the latter stages of a test or omit or randomly complete end-of-test items, these items tend to appear harder than they do when they are administered under nonspeeded conditions (Bejar, 1985; Bolt et al., 2002; Oshima). Because speededness produces noise in examinees' responses, lowering the reliability of the test. Thus, coefficient alpha is a lower bound reliability estimate in speeded tests that do not penalize guessing (Attali, 2005). This condition makes the interpretation of internal consistency measures difficult. Test-retest is the preferred reliability measure for speeded tests.

Because the HAB work samples are neither pure speeded nor pure power tests, both test-retest and internal consistency reliability measures were used in this study. Two commonly used measures of internal consistency were calculated: Chronbach's alpha (Chronbach, 1951) and Guttman's split-half (Guttman, 1945). Where appropriate, samples comprised only of examinees who attempted all test items were included in the calculation of coefficient alpha. Research supports using only the nonspeeded examinees for equating and estimating item parameters (Wallach & Wells, 2003).

Internal Consistency Reliability

Sample

The initial sample included data from all tests administered between January 1, 2007 and May 23, 2012. Cases with incomplete demographic data were removed from the initial sample – these cases accounted for less than one percent of the initial sample. The cases are fairly evenly distributed across the years except for 2012 which was an incomplete year at the time of the study. Table 1 shows the sample sizes by year.

Table 1: Frequency and Percent of Cases by Year

Year	Frequency	Percent
2007	2,457	21.0
2008	2,436	20.8
2009	2,038	17.4
2010	1,920	16.4
2011	2,802	24.0
2012	44	00.4
Total	**11,697**	**100**

A slight majority of the cases were male (66.9%), a pattern that is fairly consistent across the years.
Table 2 shows the breakdown by sex across the years.

Table 2: Frequency and Percent of Cases (Year by Sex)

Sex	2007	2008	2009	2010	2011	2012	Total
Female	994	986	856	896	1,292	17	**5,041**
	40.5%	40.5%	42.0%	46.7%	46.1%	38.6%	**43.1%**
Male	1,463	1,450	1,182	1,024	1,510	27	**6,656**
	59.5%	59.5%	58.0%	53.3%	53.9%	61.4%	**56.9%**
Total	**2,457**	**2,436**	**2,038**	**1,920**	**2,802**	**44**	**11,697**

.

The largest single contributing age group was the 40-55 year olds which comprise approximately one-third (35.9%) of the cases. Table 3 provides an age group by year breakdown.

Table 3: Frequency and Percent of Cases (Year by Age Group)

age group	2007	2008	2009	2010	2011	2012	Total
15-21	92	234	327	439	843	26	**1961**
	3.7%	9.6%	16.0%	22.9%	30.1%	59.1%	**16.8%**
22-30	444	410	380	372	526	3	**2,135**
	18.1%	16.8%	18.6%	19.4%	18.8%	6.8%	**18.3%**
31-39	512	538	434	398	502	5	**2,389**
	20.8%	22.1%	21.3%	20.7%	17.9%	11.4%	**20.4%**
40-55	1090	989	731	603	782	7	**4,202**
	44.4%	40.6%	35.9%	31.4%	27.9%	15.9%	**35.9%**
56+	319	265	166	108	149	3	**1,010**
	13.0%	10.9%	8.1%	5.6%	5.3%	6.8%	**8.6%**
TOTAL	**2,457**	**2,436**	**2,038**	**1,920**	**2,802**	**44**	**11,697**

Findings

Internal consistency reliability estimates are shown in Table 4. A general rule is that coefficients ≥ .70 demonstrate acceptable levels of reliability (Nunally & Burnstein, 1994). All of the work samples met or exceeded this standard with the exception of Concept Organization. The reliability of this work sample can be improved by adding more items. It is a relatively short test (10 questions) and short tests tend to have low internal consistency reliability.

Table 4: Internal Consistency Reliability Coefficients by Work Sample

Work Sample	N	Reliability Coefficient	
		Alpha	Split-Half
Classification	11,692	.81[a] (.70[b])	.73[a] (.70[b])
Concept Organization	11,737	.57	.55
Design Memory	11,863	.76	.71
Extravert-Introvert	11,761	.80	.79
Generalist-Specialist (written)	11,945	.88	.85
Generalist-Specialist (oral)	11,715	.88	.86
Idea Productivity	n/a	n/a	n/a
Number Memory	11,740	.83	.83
Observation	11,755	.73	.71
Pitch Discrimination	11,804	.89	.82
Rhythm Memory	11,820	.75	.73
Spatial Relations Theory	11,937	.88	.84
Spatial Relations Visualization	11,784	.74	.72
Time Frame	11,677	.93	.88
Tonal Memory	11,816	.89	.85
Typing Speed	n/a	n/a	n/a
Verbal Memory	11,889	.89	.83
Visual Speed/Accuracy	11,976	.99 (n=41[c])	.99 (n=41[c])
Vocabulary	11,633	.96	.93

[a] items not attempted were included in the data – assigned 0 points; items scored as correct/incorrect

[b] items not attempted were treated as missing; items scored with point system

[c] included only the 41 cases in which all items were attempted

Test-Retest Reliability

Sample

The sample is comprised of 95 junior/senior level psychology majors enrolled in a large southeast public university. The sample sizes for each work sample ranged between 84 and 95 (a few students failed to complete all the work samples in the battery).

Method

The students were administered the HAB under standard conditions and instructions. They completed the HAB twice with a 4-week interval between test administrations. Pearson correlations were calculated between Time 1 and Time 2 test scores for each work sample.

Findings

Table 5 shows that all work samples met or exceeded the 0.70 minimum acceptable level of test-retest reliability.

Table 5: Correlations, Means, Standard Deviations, and Sample Sizes for Time 1 and Time 2 Test Administrations

Work Sample	*r_{12}	Mean	Std	n
Visual Speed (5 min)	0.70	84.09	14.57	86
		85.20	14.94	
Visual Speed (6 min)	0.74	97.71	16.14	91
		98.13	16.13	
Visual Accuracy (5 min)	0.81	6.91	6.93	84
		6.08	6.82	
Visual Accuracy (6 min)	0.73	7.97	7.52	93
		7.07	8.86	
Typing	0.94	327.73	95.21	93
		340.25	104.35	
Generalist-Specialist (written)	0.81	6.13	4.33	86
		6.07	3.92	
Spatial Reasoning Theory	0.88	45.06	33.14	95
		53.82	38.97	
Idea Productivity	0.80	184.72	67.12	94
		175.66	64.96	
Verbal Memory	0.80	31.57	17.51	79
		37.74	18.14	
Design Memory	0.80	63.70	19.63	84
		65.79	19.73	
Tonal Memory	0.83	26.59	6.31	92
		26.74	6.76	
Rhythm Memory	0.80	31.63	4.49	91
		31.57	5.43	
Pitch Discrimination	0.86	44.63	8.91	93
		44.52	10.97	
Spatial Relations Visualization	0.71	7.19	2.96	90
		7.69	3.29	
Observation	0.76	38.78	16.92	90
		44.38	16.94	

Table 5 (continued): **Correlations, Means, Standard Deviations, and Sample Sizes for Time 1 and Time 2 Test Administrations**

Work Sample	*r_{12}	Mean	Std	n
Concept Organization	0.76	64.03	22.15	90
		67.00	23.16	
Number Memory	0.84	19.16	25.69	90
		21.43	26.34	
Time Frame	0.76	42.38	15.76	94
		43.85	15.60	
Classification	0.78	98.23	42.02	79
		106.13	48.10	
Generalist-Specialist (oral)	0.83	10.16	6.10	86
		10.58	6.25	
Vocabulary	0.90	43.39	13.83	94
		46.07	15.15	
Introversion-Extraversion	0.86	10.03	4.90	95
		9.75	4.98	

Conclusions

The HAB tests meet or exceed professionally developed standards for test reliability. HAB users can be confident that test scores are consistent over time and across items. It is recommended that the HAB reliability continued to be assessed on a regular basis using both classical test theory methodology (test-retest, internal consistency) and new item response theory (IRT) methodology where appropriate.

References

Anastasi, A. (1982). *Psychological Testing*. Fifth Edition, Macmillan Publishing Co., New York.

Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, Vol. 29, 5, 357–368.

Bejar, I. I. (1985). Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language (Report No. ETS-RR-85-11). Princeton, NJ: Educational Testing Service.

Chronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:3, 297-334.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10 (4), 255-282.

Nunnally J., & Bernstein L. (1994). *Psychometric theory*. New York: McGraw-Hill Higher, Inc.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, 31, 200-219.

References (continued)

Schnipke, D. L. and Scrams, D. J. (*1997*). Modeling item response times with a two-state mixture model: a new method of measuring speededness. *Journal of Educational Measurement* 34 (3), 213-232.

Swineford, F. (1956). *Technical manual for users of test analysis. Statistical Report 56- 42*. Princeton, NJ: Educational Testing Service.

Wollack, A., Cohen, A. and Wells, C. (2003). A Method for Maintaining Scale Stability in the Presence of Test Speededness. *Journal of Educational Measurement*, Vol. 40, 4, 307-330.