Johnson O'Connor Research Foundation, Inc.

STATISTICAL BULLETIN 2018-2

Summary of Long-Term Stability Findings

David Schroeder
February 19, 2018

[Summary, by Brian Oberlander: This SB summarizes the results of a large-scale study of long-term stability for 18 tests in the foundation's battery. The goal is to help summarizers "gauge the extent to which our obtained scores are likely to be influenced by measurement error." Note: the advice is not to provide this information directly to examinees, and instead to reserve it for each summarizer's own consideration when interpreting scores. The SB contains standard error calculations for 1] the short-term coefficient, or the potential difference between the examinee's measured and "true" scores in short-term retests; 2] the disattenuated coefficient, or the likelihood that an examinee's underlying aptitude has changed between testing and retesting; and 3] the long-term stability coefficient, which combines the previous two items to consider potential variations in both the test-retest measurements and the aptitude itself over time. Overall, this study indicates "rather high levels" of stability for the underlying aptitudes that we measure, but a lower stability for individual measurements and short-term retest scores for several of the foundation's tests (e.g. Observation and Tweezer Dexterity).]

Johnson O'Connor Research Foundation, Inc.

STATISTICAL BULLETIN 2018-2

Summary of Long-Term Stability Findings

David Schroeder
February 19, 2018

With the completion of our study of Number Facility, we have now examined the long-term stability of scores on 18 of our tests (SBs 1998-5, 2012-15, 2013-12, 2014-10, & 2018-1; TR 1997-1).[1] The purpose of this SB is to summarize the results for the 18 tests and then provide standard errors for individual scores based on the stability coefficients. These standard-error values allow summarizers to gauge the extent to which our obtained scores are likely to be influenced by measurement error. One probably would not report the standard errors directly to the examinee but rather would keep them in mind while interpreting scores.

Rationale

Each standard error was calculated via the standard formula for the standard error of measurement (Anastasi & Urbina, 1997, p. 107): SEM = SD*SQRT(1-r_{tt}). The standard error for the short-term coefficient indicates the extent to which a given examinee's score tends to differ from the examinee's "true" score (Anastasi & Urbina, 1997, pp. 107-109)--that is, from the score that the examinee would have gotten if we had taken the average of an infinite number of administrations within the short-term time frame.[2] The standard error for the disattenuated coefficient represents the extent to which an examinee's underlying aptitude is likely to have actually changed between the original testing and the retest. Finally, the standard error for the long-term stability coefficient reflects both issues: measurement error on the two administrations and change in the underlying aptitude over time.

---

[1] This discussion of stability refers to the extent to which individual differences on our tests are replicated when examinees are retested at later times—that is, whether high scorers continue to be high scorers, low scorers to be low scorers, and so on. It is distinct from the issue of whether mean scores remain the same at different ages, which we have addressed in reports on tests' age curves (e.g., SBs 2016-3 & 2016-4).

[2] In addition, it is assumed here that practice effects are controlled for and that the underlying aptitude has not changed (because of the short test-retest interval).

Results

Table 1 shows the short-term, long-term, and disattenuated coefficients from the test-retest studies of the 18 tests, along with 95% confidence intervals for the short-term and long-term coefficients.[3] For example, for Graphoria, the short-term coefficient is .85 with a 95% confidence interval of .78 to .90. Figures 1 to 3 show the coefficients in graphical form.

Table 2 shows the standard errors for individual scores in the three contexts: short-term, long-term, and disattenuated.[4] For example, for Graphoria, the standard error for an examinee's score relative to a short-term retest is 11.6 points; the standard error relative to a long-term retest is 14.6 points; and the error relative to the disattenuated coefficient (that is, the typical change in the aptitude over time) is 9.9 points.

As can be seen in Table 1, the disattenuated coefficients for our tests tend to be relatively high, with every value greater than or equal to .78. The short-term coefficients range from .56 (Observation) to .92 (Tonal Memory), and some of the values are lower than we would like to see, which is reflected in relatively large standard errors. Since the short-term samples were often small (e.g., the *N* for Observation was only 70), the confidence intervals are sometimes large, and the true values may be higher than the obtained values for tests such as Observation.

The long-term coefficients tend to parallel the short-term coefficients, which is consistent with the high values for the disattenuated coefficients. The highest long-term coefficients were for Tonal Memory (.85) and English Vocabulary (.81), while the lowest values were for Tweezer Dexterity (.53), Observation (.62), and Ideaphoria (.62).

In Table 2, as discussed, the standard errors are a function of the stability coefficients for the various tests and the corresponding standard deviations of the tests. Hence, based on the coefficients in Table 1, it follows that one sees fairly low standard errors for Tonal Memory and English Vocabulary, relative to their standard deviations, and fairly high

---

[3] Calculation of confidence intervals for disattenuated coefficients is complex, but in general the intervals are somewhat wider than the intervals for the other two coefficients because they are affected by error in both of those coefficients.

[4] These standard errors differ from the conventional standard error of measurement (SEM; SB 2012-8) in that the usual SEM represents error relative to another administration (or, rather, an infinite number of administrations) at the same time as the original administration whereas these standard errors indicate error relative to administrations at other times, with short-term or long-term time intervals between administrations. In general, conventional SEMs are smaller than standard errors across occasions.

standard errors for Observation and Tweezer Dexterity, in relation to their standard deviations.

Summary

In this report, the results of test-retest studies on 18 of the Foundation's tests are presented, along with confidence intervals for the short-term and long-term coefficients and standard errors for individual scores. In general, our tests show rather high levels of stability in the underlying aptitudes being measured. The levels of short-term stability are lower than we would prefer for several of our tests, and so we need to be aware of the degree of error in one-time measurements of those aptitudes.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Statistical Bulletin 1998-5. *Long-term stability of Finger Dexterity, Tweezer Dexterity, and Tonal Memory*. L. L. Meyer & D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 2012-8. *Standard errors of measurement for the Foundation's standard battery of tests*. D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 2012-15. *Long-term stability for Foresight.* D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 2013-12. *Long-term stability for Pitch Discrimination and Rhythm Memory.* D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 2014-10. *Long-term stability for English Vocabulary*. D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 2016-3. *Age curve for the Analytical Reasoning test*. D. H. Schroeder & L. S. Houser-Marko. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 2016-4. *Age curve for the Number Facility test*. D. H. Schroeder & L. S. Houser-Marko. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 2018-1. *Long-term stability for Number Facility*. D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Technical Report 1997-1. *Long-term stability of 11 aptitude tests*. J. K. Bethscheider & D. H. Schroeder. Chicago: Johnson O'Connor Research Foundation.

Table 1
Stability Coefficients for 18 Foundation Tests

Test	Short-term coef.	Long-term coef.	Disattenuated coef.	Source for coefficients
Graphoria	85 (78,90)	76 (71,80)	89	TR 1997-1
Ideaphoria	71 (65,76)	62 (58,66)	87	TR 1997-1
Foresight	76 (68,82)	64 (56,71)	84	SB 2012-15
Inductive Reasoning	67 (57,75)	64 (58,69)	96	TR 1997-1
Analytical Reas.	65 (44,79)	63 (55,70)	97	TR 1997-1
Number Facility	68 (58,76)	65 (58,71)	96	SB 2018-1
Wiggly Block	82 (73,88)	65 (57,72)	79	TR 1997-1
Tonal Memory	92 (88,95)	85 (82,87)	92	SB 1998-5
Pitch Discrimination	88 (81,93)	75 (71,79)	85	SB 2013-12

(*table continues*)

Table 1 (*continued*)

Test	Short-term coef.	Long-term coef.	Disattenuated coef.	Source for coefficients
Rhythm Memory	68 (57,77)	69 (64,74)	100[a]	SB 2013-12
Memory for Design	77 (65,85)	73 (66,79)	95	TR 1997-1
Silograms	81 (70,88)	73 (66,79)	90	TR 1997-1
Number Memory	73 (59,83)	69 (61,76)	95	TR 1997-1
Observation	56 (38,70)	62 (53,69)	100[a]	TR 1997-1
Finger Dexterity	77 (65,85)	64 (56,71)	83	SB 1998-5
Tweezer Dexterity	66 (55,75)	53 (46,60)	80	SB 1998-5
Word Association	81 (73,87)	63 (57,68)	78	TR 1997-1
English Vocabulary	89 (82,94)	81 (77,84)	92	SB 2014-10

Note. For the short-term and long-term coefficients, 95% confidence intervals are shown on the lines below the coefficients. Decimals are omitted in all values.

[a] For Rhythm Memory and Observation, the long-term coefficients were greater than the short-term coefficients, and so the calculated values for the disattenuated coefficients were greater than 1. Since correlation coefficients cannot be that high, those two values were set to the ceiling value for correlations, namely, 1.00 (SB 2013-12; TR 1997-1).

Table 2
Standard Errors for Individual Scores for 18 Foundation Tests

Test	SE for short-term retest	SE for long-term retest	SE for disat-ten. retest	Standard deviation[a]
Graphoria	11.6	14.6	9.9	29.9
Ideaphoria	35.2	40.3	23.6	65.4
Foresight	10.1	12.4	8.3	20.7
Inductive Reasoning	12.8	13.4	4.5	22.4
Analytical Reasoning	7.6	7.9	2.2	12.9
Number Facility	10.0	10.4	3.7	17.6
Wiggly Block	30.9	43.1	33.4	72.8
Tonal Memory	3.2	4.4	3.2	11.3
Pitch Discrimination	3.5	5.1	3.9	10.2
Rhythm Memory	2.8	2.7	--	4.9
Memory for Design	12.8	13.9	6.0	26.8
Silograms	4.2	5.0	3.0	9.6
Number Memory	15.0	16.1	6.5	28.9
Observation	8.8	8.2	--	13.3
Finger Dexterity	5.7	7.1	4.9	11.8

(*table continues*)

Table 2 (*continued*)

Test	SE for short-term retest	SE for long-term retest	SE for disat-ten. retest	Standard deviation[a]
Tweezer Dexterity	10.3	12.1	7.9	17.7
Word Association	3.4	4.7	3.7	7.8
English Vocabulary	9.7	12.7	8.2	29.1

[a] The standard deviations shown in this column, which were used in the calculation of the standard errors, were derived from the Foundation's database of raw scores for code-0 examinees from 2011 to 2015, with age effects statistically removed. For Inductive Reasoning, only scores for Form OA were used in the calculation of the standard deviation. For Wiggly Block, only scores for Form EN were used.

Figure 1
Short-Term Stability Coefficients for 18 Foundation Tests

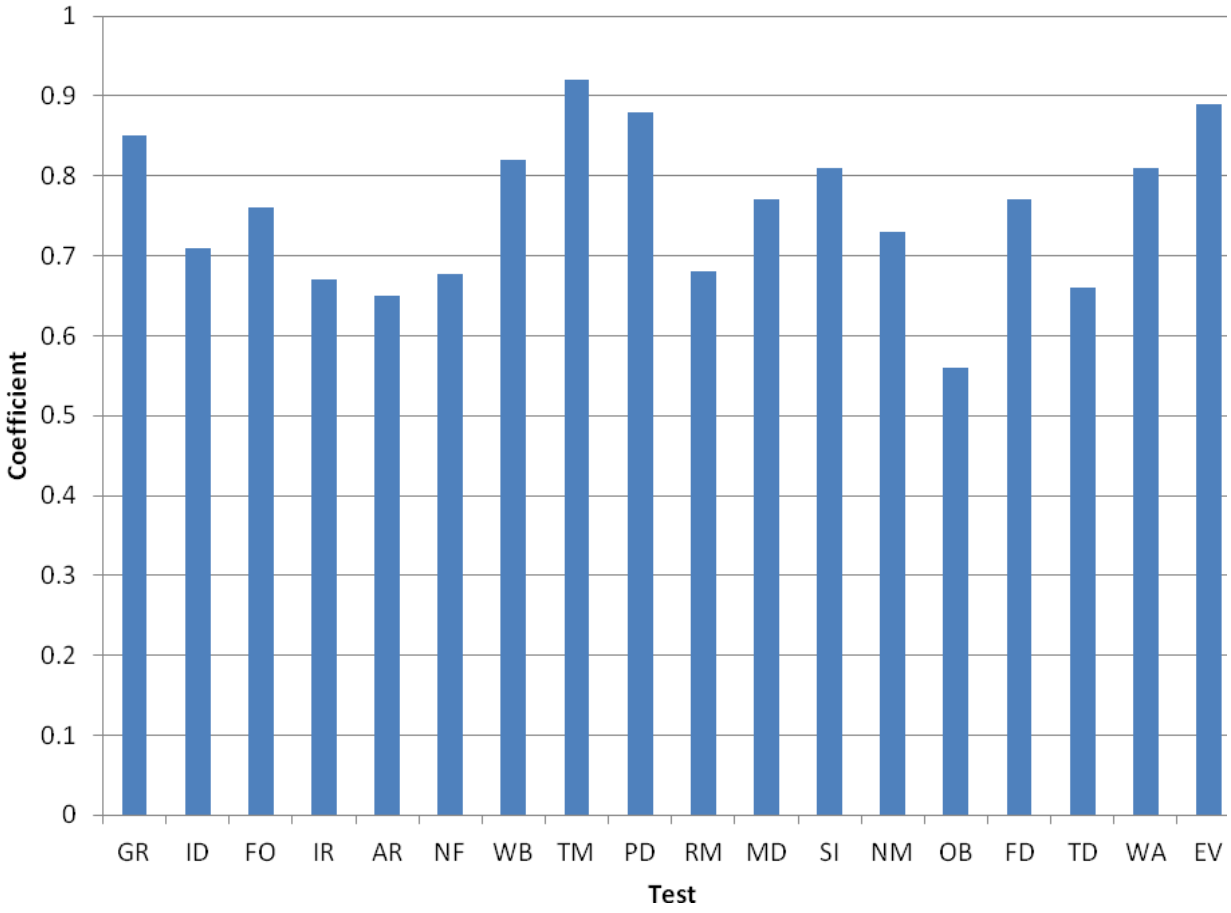

Figure 2
Long-Term Stability Coefficients for 18 Foundation Tests

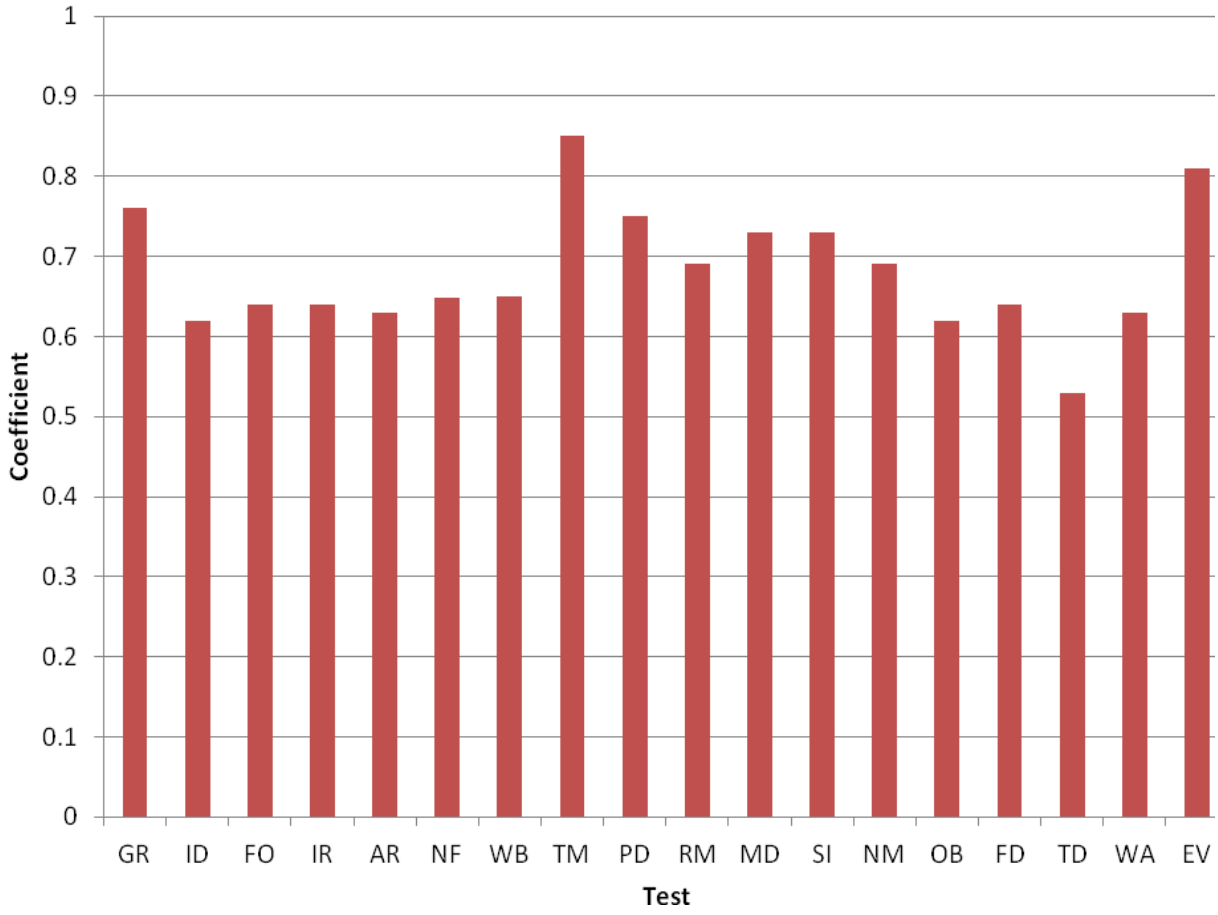

Figure 3
Disattenuated Stability Coefficients for 18 Foundation Tests