Cognitively SPEAKING

An update on Ability Measurement with *CogAT*®
Volume #1 • Summer Issue • 2004

Riverside Publishing
A HOUGHTON MIFFLIN COMPANY

Welcome to the first issue of *Cognitively Speaking.*

Each issue will address a topic that is relevant for *Cognitive Abilities Test*™ (*CogAT*®) users. If you would like to submit a topic, please go to www.cogat.com and use the online Feedback Form. We appreciate your feedback!

This first issue of *Cognitively Speaking* discusses a question that perplexes many test users: Why do the same student's scores on different ability tests often disagree?

One teacher of gifted students writes:

Our school psychologist frequently uses the *Wechsler Intelligence Scale for Children*® *(WISC*®*)* when evaluating students. I have administered *CogAT* to some of the same students. Even if the *WISC* score is "very superior" (e.g., a Full Scale Score of 133), the individual's *CogAT* scores are usually lower. In some instances, the student previously took the *Otis-Lennon School Ability Test*® *(OLSAT*®*)* or the *Stanford-Binet Intelligence Scales,* which aligns more closely with the *CogAT* scores. What is going on?

A parent asks a similar question:

What is the difference between *CogAT* and the *WISC?* Last year my son's Composite national percentile rank was 88 on *CogAT*. This year he scored at the 99th percentile on the *WISC*. Is this normal?

How much agreement should we expect between scores on *CogAT* and those on other ability tests? Test developers typically report that scores on ability tests are *highly correlated* (r = .70 or higher). In a recent study, the *CogAT* authors found that the Composite score on *CogAT* Form 6 correlated r = .69 with the General Ability Index on the *Woodcock-Johnson*® *III (WJ III*®*) Tests of Cognitive Abilities.* In another study, the Composite score on *CogAT* Form 6 correlated r = .79 with the *WISC-III* Full Scale Score. (See reference #1 under "For More Information.")

Test developers claim that correlations between ability test scores are generally high. Yet, we know from experience that a student's scores on different ability tests frequently disagree, even when the items on the different tests define the same ability factor. Suppose the correlation between two ability tests is r = .80. Statistical analyses show that of the students who score in, say, the percentile range of 80–90 on the first test, *only 26.1* percent will also score in the percentile range of 80–90 on the second test. High correlations imply much less agreement than we would expect. (See reference #2 under "For More Information.") The parent who asked about the discrepancy between his son's *CogAT* and *WISC* scores can be reassured that this discrepancy is not uncommon.

The teacher who works with gifted students may see a somewhat different problem. Schools commonly use achievement test results to screen students for admission to special programs. Individuals who meet the selection criterion of, say, the 90th percentile on the achievement test *also* take an ability test. If the two tests correlate r = .80, *only 56 percent* of the students scoring above the 90th percentile on the achievement test will also score above the 90th percentile on the ability test. The other 44 percent score below the 90th percentile. This means that the average score on the ability test for these students will be lower than the average score on the achievement test.

CogAT®

In general, then, when an achievement test is used to screen students, scores of the high-scoring students tend to be lower on the ability test. This is one reason many experts recommend screening *all* students with an achievement test *and* an ability test. Administering both tests to all students will identify individuals who score higher on the ability test than on the achievement test. These individuals are missed when only students who pass the cut score on the screening test can take the second test.

Why do scores on two ability tests differ? The most important reasons are:

- **Different tests measure different aspects of the same ability.**
- **Each test is normed on a different group of students.**
- **Errors of measurement occur.**

Different tests measure different aspects of the same ability. There is no single indicator of ability, just as there is no single indicator of health. For this reason, different ability tests present different tasks, and, therefore, estimate somewhat different aspects of the same ability construct. For example, the *Woodcock-Johnson III Tests of Cognitive Abilities* measures eight broad ability factors including reasoning, phonemic awareness, decision speed, short-term memory, and verbal information. *CogAT,* on the other hand, measures the ability to reason in three symbol systems (verbal, quantitative, and nonverbal). By design, then, *CogAT* measures somewhat different abilities than the *Woodcock-Johnson III.*

Different task formats also make a difference. On some tests, students read the items to themselves. On others, the test administrator reads the items aloud. If a student has serious reading difficulties, even short sentences pose problems. For this reason, *CogAT* is orally administered at grades K–2. However, some English Language Learners will do better on *CogAT* if administered a level that they can read themselves. For these students, understanding the teacher may be more difficult than reading the items.

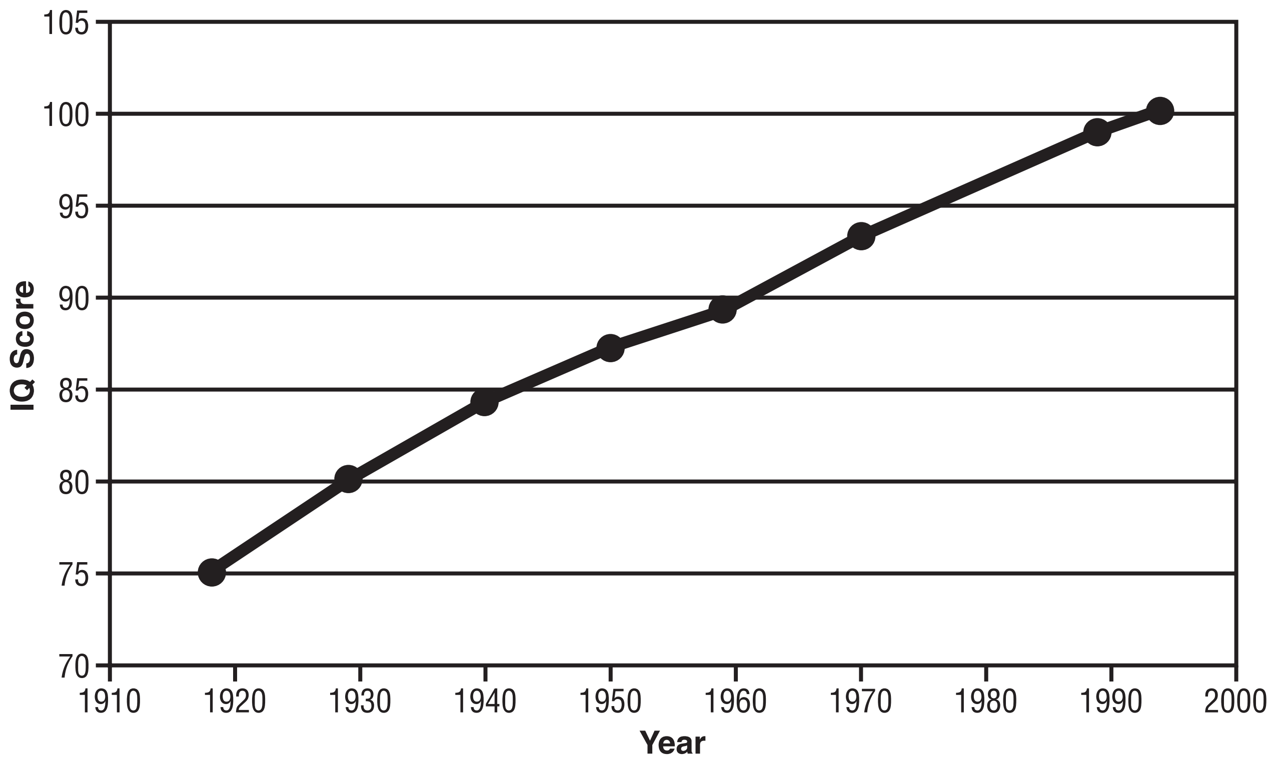
The important idea here is that format effects can be large and can differ across students. When using a test to predict academic success, consider how well the format duplicates the manner in which students will typically learn new concepts and skills. For very young children, the primary means of classroom learning are the teacher's oral explanations. As students mature, written texts become increasingly important. Even so, reading skills required on *CogAT* are minimal compared to a reading comprehension test at the same grade level.

Each test is normed on a different group of students. Norming samples can vary substantially, especially if the number of students is small. Typically, norming samples for group-administered tests contain only children who attend school. Children who do not attend school and school children who are unable to take the test, even with accommodations, are not included. Developers of individually administered intelligence tests, however, try to include all of these children in their norming sample.

Norms change over time. Figure 1 shows one estimate of the growth in average scores on the *Wechsler* and *Binet* IQ tests during the twentieth century. A student who earned a score of 100 in 1920 would have received a score of 75 in 1996. However, the mean score remains 100 because tests are continually renormed to make it so.

Figure 1

Flynn Effect Chart



Gains in Wechsler-Binet IQ for the U.S. White population.
Sources J. Horgan (1995) and D. Schildlovsky.

The implication is that students who take tests normed at different points in time can receive different scores for the same performance. For example, if we compare the norms for *CogAT* 6 with those for *CogAT* 5, we will see that a nine-year-old who scored at the 69th percentile in 1992 would have scored at the 50th percentile in 2000 for the same performance. (See reference #3 under "For More Information.") This is probably due to changes in the mathematics curriculum. The emphasis in primary-grade mathematics instruction has changed from learning basic computation skills to developing quantitative reasoning abilities. Abilities respond to instruction, especially abilities that are not emphasized outside of school.

Errors of measurement occur. Every test presents a particular sample of items on a particular occasion. Administering a second sample on a second occasion gives a somewhat different estimate of ability for most students. We call differences between the two estimates *errors of measurement.* If tests are reliable, these errors are small for most test takers. Sometimes errors of measurement for a particular student are larger due to inconsistencies in the way the individual responded to different item types or to different subtests. If this occurs on *CogAT* 6, the confidence interval around the affected battery score will be very wide, indicating confusion on the part of the student and signaling a battery score that should be used with caution.

Errors of measurement on scale scores, such as Standard Age Scores, increase dramatically as the individual's raw score approaches the highest possible score on the test. (See reference #5 under "For More Information.") For this reason, the *CogAT* authors strongly recommend testing very able students with a higher test level than is generally recommended for the student's grade. Part 1 of the *CogAT Form 6 Research Handbook* provides specific suggestions for selecting the appropriate test level for very able students.

Detecting systematic errors that affect the entire test session is not easy. For example, illness, anxiety, or lack of motivation on the day of testing can lower scores. The only way to detect such errors is to retest the student. Repeated administrations using the same form of a test can raise scores because of "practice effects," especially if the retest comes soon after the initial administration. An alternative solution for *CogAT* 6 is to retest with *CogAT* 5 and then use tables from the *CogAT* 5 / *CogAT* 6 Equating Study to estimate the corresponding score on the *CogAT* 6 scale. These tables are available from Riverside Publishing.

Summary

The expectation that a student will obtain similar scores on different ability tests stems from a misunderstanding of correlations. Only a minority of students obtain similar scores on two different, but highly correlated, ability tests. This is a statistical fact of life.

Scores on two ability tests administered to the same student differ primarily because the ability tests measure somewhat different aspects of the same ability construct. For example, *CogAT* measures abstract reasoning abilities in the three symbol systems most required for learning in school: verbal, quantitative, and figural/spatial.

Other tests, especially those used for clinical work, try to sample a much broader array of specific and general abilities. This does not make them better predictors of school success; indeed, *CogAT* generally predicts school success better than individually administered ability tests. Clinical tests are designed to provide insight into specific cognitive functions that can suggest the source of learning difficulties. Furthermore, the various individually administered ability tests define the domain of intelligence somewhat differently, and scores on these tests correlate with each other about as well as they correlate with *CogAT*.

Recommendations

1. **Expect scores to differ across tests, even when the tests are highly correlated.** See the tables on prediction efficiencies in reference #2 of "For More Information" to learn how much variation to expect with different correlations.

2. **Do not attribute discrepancies between tests to flaws in one of the tests.** The most common mistake here is to accept one score, especially if it is from an individually administered test, and to discount the other, especially if it is from a group-administered test. Tests measure different abilities and have different purposes.

3. **Use tests with recent norms.** Discrepancies between tests are more likely if the tests are normed at different times or if the norming samples are dissimilar.

4. **Look beyond total scores when attempting to understand large score discrepancies.** On *CogAT* Form 6, look at the number of items attempted, at the width of the confidence intervals around the scores, and at any score warnings reported for the student. Keep in mind that the *CogAT* Composite score is most meaningful when the Verbal, Quantitative, and Nonverbal Battery scores are all at about the same level (an "even" or "A" profile). Similarly, look at the extent to which subtest scores on an individually administered test either scatter or agree with one another. The Composite score on any test is less dependable when scores on parts of the test display unusual scatter.

For More Information

1. For validity studies comparing *CogAT* Form 6 with the *Woodcock-Johnson III* and the *Wechsler Intelligence Scale for Children III,* see http://faculty.education.uiowa.edu/dlohman.

2. For examples of how to interpret correlations in terms of the precision of the prediction from one test to another, see http://faculty.education.uiowa.edu/dlohman/pdf/tables_of_prediction_efficiencies.pdf. For a comparison of *CogAT* 5 Standard Age Scores (1992 norms) with those of *CogAT* 6 (2000 norms), see Tables 6.2–6.5 in the *CogAT Form 6 Research Handbook.*

3. For information on increases in scores on ability tests over the past hundred years, see U. Neisser (1998). *The rising curve: Long term gains in IQ and related measures.* Washington, DC: American Psychological Association.

4. Figure 1 is adapted from J. Horgan (1995). *Scientific American,* 273 (5), p. 14 reproduced in M. Martinez (2000). *Education as the cultivation of intelligence.* Mahwah, NJ: Erlbaum.

5. For more information on how errors of measurement increase as scores approach the maximum number correct, see the section titled "Conditional Standard Errors of Measurement" in Part 5 of the *CogAT Form 6 Research Handbook.*

Riverside Publishing
A HOUGHTON MIFFLIN COMPANY

Cognitively Speaking **is a publication of Riverside Publishing.**

Riverside Publishing
425 Spring Lake Drive
Itasca, IL 60143-2079

For more information, call Customer Service at 800.323.9540 or visit www.riversidepublishing.com and www.cogat.com.