Cognitively SPEAKING

An update on Ability Measurement with *CogAT*®
Volume #6 • WINTER Issue • 2008

Riverside Publishing
A HOUGHTON MIFFLIN COMPANY

Comparing CogAT, NNAT, and the Raven

by David F. Lohman, The University of Iowa

School personnel who administer more than one ability test to students are frequently puzzled when an individual's test results disagree. In the first edition of *Cognitively Speaking* (see www.cogat.com), I discussed three reasons for this: (1) differences in the abilities tested, (2) errors of measurement, and (3) regression to the mean (the tendency for an individual's extreme scores on one test to be less extreme on a second, related test). In this edition of *Cognitively Speaking*, I discuss a fourth reason: *differences in the quality of test norms.*

Differences in the quality of test norms often result in systematic differences in the number of students whose scores fall above (or below) a particular value. Some inadequacies in norms are obvious; however, others can be difficult to document. Even careful scrutiny of test manuals by educators trained in psychometrics may not uncover facts that affect the accuracy of the norms. Consequently, school personnel may not realize they are using tests with defective norms and scores that can not be trusted. The *Culture-Fair Intelligence Test* (*CFIT*; Cattell & Cattell, 1965) is an example of a test with obviously defective norms. Even though *CFIT*'s norms were woefully inadequate when the test was first published (Tannebaum, 1965), the norms have never been replaced. Not surprisingly, in 2004 Shaunessy et al. noted that IQ scores on the *CFIT* are, on average, 17.8 points higher than those on the *Naglieri Nonverbal Abilities Test*® (*NNAT*®). Nonetheless, the test is often recommended to assess the ability of low-socioeconomic-status (low-SES) students and minority students (e.g., Mulligan, 2007).

Over the years, the authors of the *Cognitive Abilities Test*™ (*CogAT*®) have conducted research studies comparing *CogAT* to carefully normed, individually administered ability tests. Two recent studies compared *CogAT* with two of the best individually administered ability tests: the *Woodcock-Johnson*® *III* and the *Wechsler Intelligence Scales for Children*®, *Third Edition* (*WISC*®). Both studies found congruence in the normative scores. For example, in the second study, the average *CogAT* Composite was identical to the average *WISC III* Full Scale IQ score. Importantly, when placed on a common scale, the variability of scores on the two tests was also the same.[1] Such congruence of both the *mean* (location of the average score) and the *standard deviation* (degree of spread, or dispersion, of scores) bolsters confidence in the normative scores of both tests.

Differences in the quality of test norms can result in differences in the number of students who obtain exceptionally high or low scores.

Comparison studies with group-administered ability tests are also important. The studies summarized in this newsletter show how the *CogAT* Nonverbal Battery fares when compared to two other nonverbal tests: the *Naglieri Nonverbal Abilities Test* and J. C. Raven's *Standard Progressive Matrices.*

CogAT®

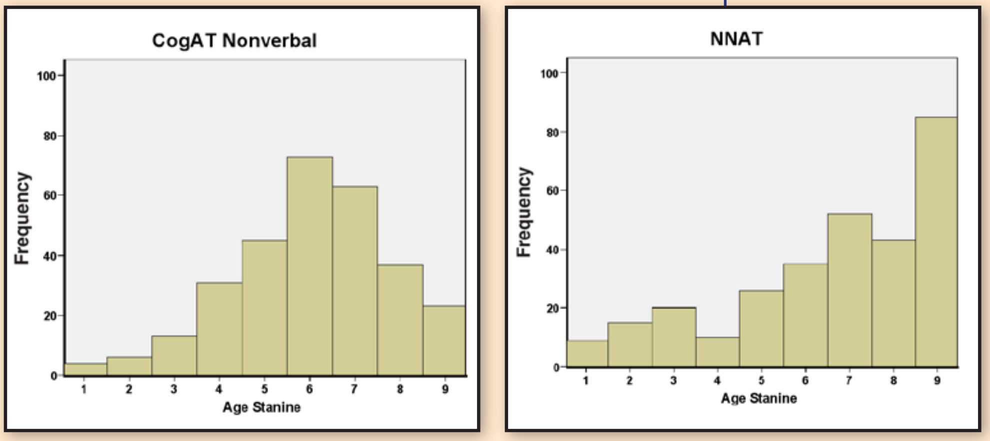
[1] For more information, see the section "Concurrent Prediction of Ability" in the *CogAT Form 6 Research Handbook.* For details on the studies comparing *CogAT* 6 with the *WISC-III* and the *Woodcock-Johnson III*, see http://faculty.education.uiowa.edu/dlohman.

Study 1: Comparing CogAT with NNAT

In Study 1, 295 kindergarteners in a high-SES Midwestern school district took both *CogAT* and *NNAT*. As would be expected in a high-SES district, the scores on both tests were above average. However, the standard deviation (SD) for the *NNAT* scores (19.0) was much greater than the SD for the *CogAT* Nonverbal Battery scores (14.3). This is surprising. If anything, the SD for the *NNAT* scores should be somewhat smaller than the SD for the *CogAT* scores. *NNAT* should have a population SD of 15, whereas *CogAT* was standardized to have population SD of 16. The large SD on *NNAT* (19.0) means that more students in this sample obtained unusually high or low scores on *NNAT* than the norms tables lead users to expect.

Pictures of the distributions of scores for the two tests show more clearly what happened. As Figure 1 shows, the distribution of *NNAT* scores was markedly skewed. In a high-SES school, it is not unusual for 7-8 percent of the students to obtain stanine scores of 9 (the highest stanine score). This is what *CogAT* showed. On *NNAT*, however, 85 students (29 percent) obtained stanine scores of 9. Note also that relatively more students received very low scores (stanines of 1, 2, or 3) on *NNAT* than on the Nonverbal Battery of *CogAT*. This suggests that something may be seriously amiss with the norms for Level A of *NNAT*.

Figure 1. Distributions of Age Stanine Scores for 295 Kindergarten Students on the Nonverbal Battery of *CogAT* 6 (left panel) and on *NNAT* (right panel)



Study 2: Comparing CogAT, NNAT, and the Raven

Conducted in an elementary school in the Southwest, Study 2 (Laing & Castellano, 2007[2]) compared the use of the *Raven*™ *Standard Progressive Matrices*, *NNAT*, and *CogAT*[3] to identify academically gifted ELL students. Most of the students were eligible for free or reduced-priced lunch. Examiners used English or Spanish directions, as appropriate.[4]

The three tests were administered in a *counterbalanced* order. This means that approximately one-third of the students were tested first with the *Raven*, one-third with *NNAT*, and one-third with *CogAT*. Similarly, one-third were administered each test in the second position and one-third in the third position. Counterbalancing eliminates the possibility that students might perform better (or worse) on one test because it is always administered first (or last).

The analyses showed that the three tests differ importantly in the quality of their norms and in their ability to identify the most academically able ELL and English-proficient Hispanic students. A summary follows.

***Raven* Scores.** When placed on the same scale as *CogAT* (mean = 100, SD = 16), scores for both ELL and English-proficient students were 10 to11 points higher on the *Raven* than on either the *CogAT* Nonverbal Battery or *NNAT*. This means that the 1986 U.S. norms for the *Raven* are markedly easier than the 2000 *CogAT* norms and the 1995–1996 *NNAT* norms. In part, this is because the 1986 *Raven* norms are not based on a representative U.S. sample but on a compilation of test scores from school districts that submitted their scores to the test authors over the years. Also, scores on nonverbal tests have risen dramatically over the past 30 years. Even if the *Raven* norms were reasonably accurate in the 1970s when the data were collected, they are far too easy today.

[2] Dr. Naglieri and I were contributing partners to this study, and were given the data to analyze. This summary is based on the analyses that Katrina Korb, Joni Lakin, and I performed on the data. For a copy of this report, see Lohman, D. F., Korb, K., & Lakin, J. (in press).

[3] Although all three *CogAT* batteries were administered to most of the children in this study, it is perfectly acceptable for schools to administer only one battery or any combination of two batteries. However, proper interpretation of scores for ELL students taking the Verbal and Quantitative batteries requires using more focused comparison groups in addition to the national (or local) norms. When this is not possible, only the *CogAT* Nonverbal Battery is typically administered.

[4] Spanish directions for *CogAT* may be obtained from Riverside Publishing's Customer Service Department.

***NNAT* and *CogAT* Scores.** For English-proficient students, the mean *NNAT* score was the same as the mean *CogAT* Nonverbal Battery score (100.7). For ELL students, the mean *NNAT* score was approximately two points *lower* than the mean *CogAT* Nonverbal Battery SAS score. The differences were particularly large for first- and second-grade ELL students. *CogAT* scores were, on average, seven points higher than *NNAT* scores at grade 1 and five points higher at grade 2.

As in Study 1, the scores on *NNAT* were much more variable than the scores on either *CogAT* or the *Raven*, especially from kindergarten through grade 2. In an attempt to understand the large variability of *NNAT* scores, we first checked whether the error of measurement in students' scores was larger for ELL students than for English-proficient students. Although the standard error of measurement was twice as large on *NNAT* (6.6) as on either the *Raven* (3.0) or the *CogAT* Nonverbal Battery (3.2), it was not larger for ELL students. Next, we examined the distributions of scores on all three tests. Figure 2 shows the distributions of scores on the *Raven*, the *CogAT* Nonverbal Battery, and *NNAT* for ELL students from kindergarten through grade 3. Figure 3 shows score distributions for English-proficient students.

All of these distributions should approximate a bell-shaped curve. However, most distributions for *NNAT* and the *Raven* are not bell-shaped. Very low scores on *NNAT* were much more common than expected for ELL students from kindergarten to grade 2 and for English-proficient students at grade 1. For example, the most common *NNAT* score for ELL students in grade 1 was a stanine of 1, the lowest possible stanine (see Figure 2). On the other hand, there were more high scores than would be expected on *NNAT* for English-proficient students as well. The twin problems of too many low scores for ELL students and

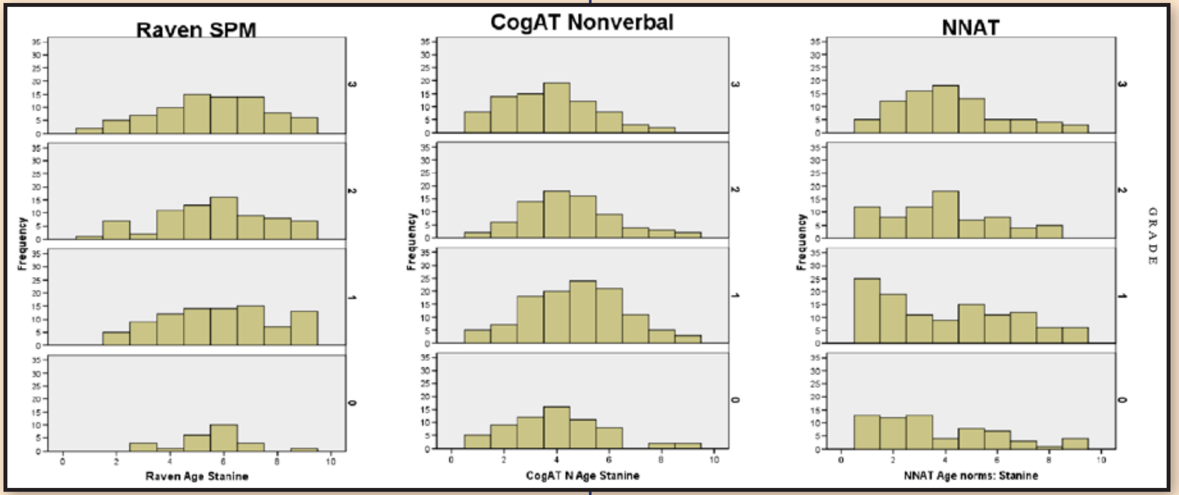


Figure 2. Distributions of Stanine Scores for English-Language Learners

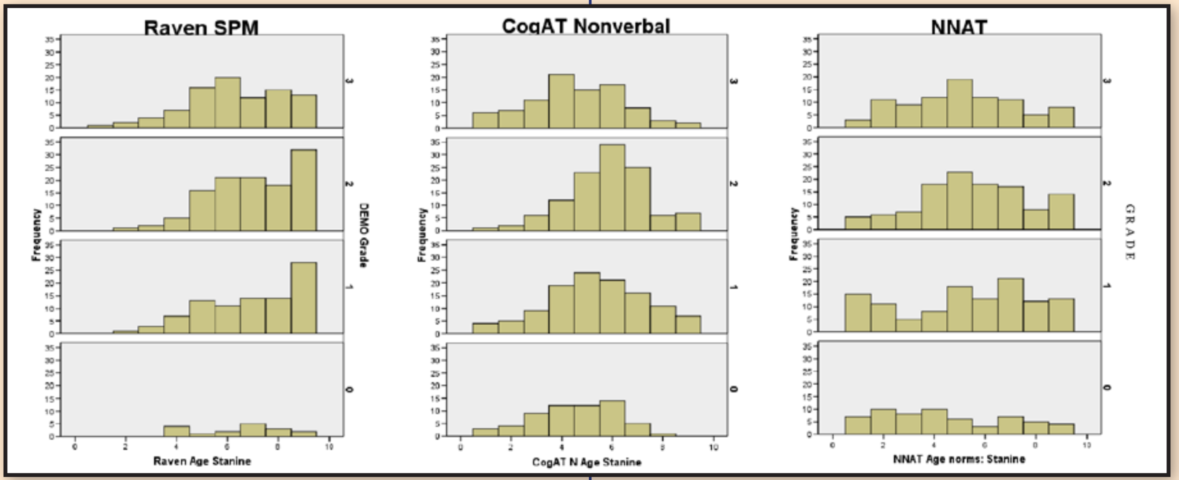


Figure 3: Distributions of Stanine Scores for English-Proficient Students

too many high scores for English-proficient students is reflected in the relatively flat distributions of *NNAT* scores in Figure 3 (see p. 3). Such distributions are consistent with the unexpectedly large SD for *NNAT*.

Comparing ELL and English-Proficient Students. Some claim that a nonverbal test provides a culture-fair measure of ability that is independent of language, background, or experience. Study 2 did not support this view. On the *Raven* and on the *CogAT* Nonverbal Battery, ELL students scored eight points lower than English-proficient students, and on *NNAT*, ELL students scored 10 points lower than English-proficient students. Further, as shown in Figure 4, the relative proportions of ELL and English-proficient students at each stanine varied dramatically across the three tests. On *NNAT*, ELL students were much more likely to receive very low scores. At the other extreme, twice as many English-proficient students received stanine scores of 9 than would be expected, given the mean Nonverbal Ability Index (NAI) score of 100.7. Only the *CogAT* Nonverbal Battery showed normally distributed scores for both student groups. Further, 3.5 percent of the English-proficient students obtained stanine scores of 9 on *CogAT*. This is exactly the proportion that would be expected, given the mean SAS score of 100.7.

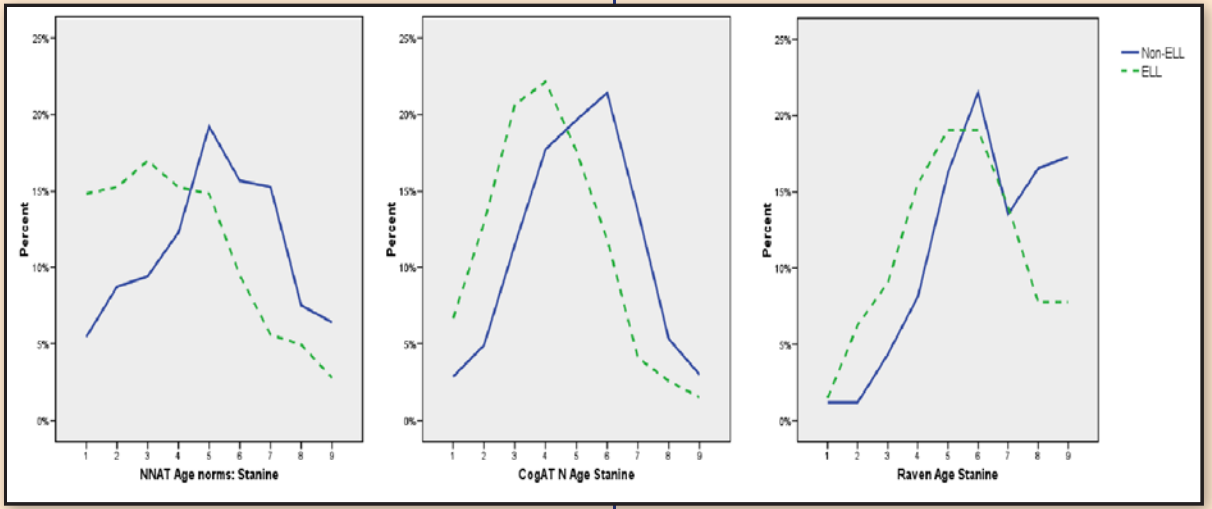


Figure 4. Percent of ELL Students (dashed line) and English-Proficient Students (solid line) at Each Stanine for *NNAT* (left panel), *CogAT* 6 Nonverbal Battery (middle panel), and the *Raven* (right panel)

Some ELL students spoke languages other than Spanish. We wondered if the differences between ELL and English-proficient students would be smaller if we included only Hispanic students who were eligible for free or reduced-price lunch. When placed on the *CogAT* SAS scale, English-proficient Hispanic students, with their increased exposure to the U.S. culture and educational system, scored higher than ELL Hispanic students by 7.5, 7.3, and 10.1 points, on the *Raven*, *CogAT*, and *NNAT*, respectively. These differences, which were generally consistent across grades, support the long-established conclusion that the impact of culture, education, and language development is reduced but not eliminated on nonverbal tests (Anastasi & Urbina, 1997). The results are also at odds with a previous report of only 1 point difference on *NNAT* between ELL and English-proficient students (Naglieri, Booth, & Winsler, 2004).

Mistakes in Norming the NNAT

That the norms for the *Raven* are untrustworthy is not surprising. The test is old and has never been properly normed. That the norms for *NNAT* could be as skewed as Study 2 suggests is surprising, especially since the test was recently normed and has been widely used for several years. Therefore, we looked for other published reports that might show the same broad dispersion of *NNAT* NAI scores observed in Studies 1 and 2. Two of the most important reports that we discovered used the original *NNAT* standardization data. George (2001) re-analyzed the Spring *NNAT* standardization data for her doctoral dissertation. In addition to finding large differences between the mean scores of Caucasian, Black, and Hispanic students[5], she reported standard deviations for "number correct" scores at each level. We used these to estimate SDs of NAI scores.

The second study was reported by Naglieri and Ronning (2000). They used the Fall *NNAT* standardization data to explore correlations between *NNAT* and the *Stanford Achievement Tests*. However, one table in their report included SDs for *NNAT*.

These two sets of standard deviations for the *NNAT* standardization, together with the SDs from the Project Bright Horizon study, are plotted in Figure 5. All three data sets show the same pattern of standard deviations. If the test had been properly normed, then all of these values would be approximately 15.

The best measure of the standard deviation at each test level is given by the Naglieri and Ronning (2000) study. Here, it is only at Level E that the SD is 15. Apparently, only the distributions of Level E NAI scores were set to a standard deviation of 15. It must have been assumed that this would also fix the SD at 15 for the other six test levels.[6]

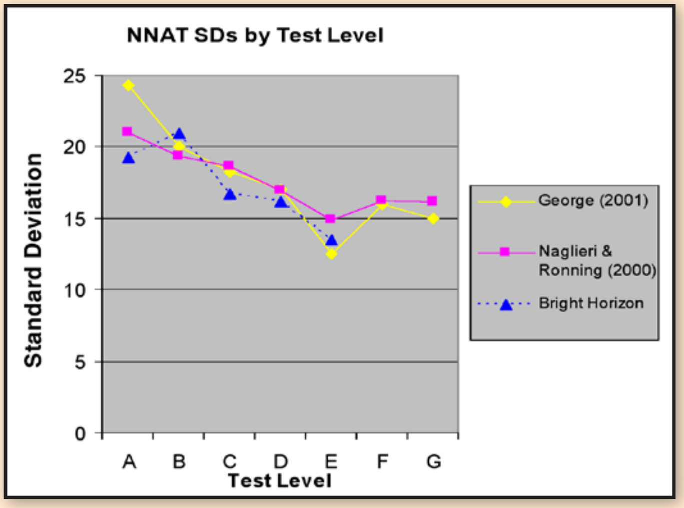


Figure 5. Standard Deviations for *NNAT* from (1) George, (2) Naglieri and Ronning, and (3) the Project Bright Horizon Study

Table 1 shows the consequences for test scores. The value at the head of each column shows the true NAI score (i.e., the value that would be observed if the SD were set to 15) for scores at the mean and then at 1, 2, and 3 SDs above the mean. The table entries show the NAI scores that are actually observed at each level of the test. For example, the student who receives an NAI score of 142 on Level A should actually have received a score of 130.

	True NAI Score			
Level	**100**	**115**	**130**	**145**
A	100	121	142	163
B	100	119	139	158
C	100	119	137	156
D	100	117	134	151
E	100	115	130	145
F	100	116	132	149
G	100	116	132	148

Table 1. True Versus Reported NAI Scores by *NNAT* Test Level

	True NAI Score		
Level	**115**	**130**	**145**
A	1.5	3.4	11.9
B	1.4	2.6	7.3
C	1.3	2.3	5.8
D	1.2	1.7	2.9
E	1.0	1.0	1.0
F	1.1	1.4	2.0
G	1.1	1.4	1.9

Table 2. Over-identification Rates for the Number of Students with NAI Scores Above 115, 130, and 145

There are several things to note in this table. First, changes in the standard deviation do not alter the mean scores. All mean scores are 100, as expected. Second, the scores at Level E correspond to the true NAI scores. This is the only level at which the SD is 15. Third, the discrepancies between reported NAI scores and true NAI scores are largest for the youngest children and the highest scores. Therefore, overestimation of the proportion of students with high scores is greatest at Level A.

[5] On the NAI scale, the median Caucasian-Black difference was 12 points. The largest difference was on level A (17 points). The median Caucasian-Hispanic difference was 5 NAI points. Again, the largest difference was on Level A (9 points).

[6] On *CogAT*, distributions of scale scores were equated not just by test level, but at 19 different percentile values for every 3-month interval from 4 years 8 months to 18+ years.

The extent to which the test over-identifies the number of high-scoring students is shown in Table 2 (see p. 5). For example, the number of students who receive NAI scores of 130 or higher on Level A is 3.4 times greater than it should be. Concretely, when both *NNAT* and a test with good norms are administered to a group of children, *NNAT* will appear to identify twice or three times as many gifted children as the properly normed test.

CogAT Verbal and Quantitative Batteries

Sometimes even good national norms are not the most appropriate reference group. This is commonly the case when one hopes to make inferences about talent (or aptitude) but students' opportunities to develop the abilities measured by the test differ markedly from those of students in the national norm group. This is the case for ELL students on the *CogAT* Verbal Battery and, to a lesser extent, on the *CogAT* Quantitative Battery. As would be expected, scores of ELL and English-proficient students were much closer to each other on the *Raven*, the *NNAT*, and the *CogAT* Nonverbal Battery than they were on the *CogAT* Verbal and Quantitative batteries. Across grades, differences between scores of ELL students and of English-proficient students were exactly twice as large on the *CogAT* Verbal Battery (16.6 points) as on the *CogAT* Nonverbal Battery (8.3 points).

Studies 1 and 2 have several implications for educators. The first is that it is wrong to assume that nonverbal tests level the playing field for children who come from different cultures or who had different educational opportunities.

Does this mean that the Verbal and Quantitative batteries are biased and that the *CogAT* Nonverbal Battery provides a better measure of academic aptitude? Not at all. In Study 2, the *CogAT* Verbal Battery scores were the best predictors of success in reading for all participants. Similarly, a weighted combination of all three *CogAT* batteries was the best predictor of mathematics achievement. Clearly, the problem is not that the *CogAT* Verbal and Quantitative batteries measure the wrong abilities. Rather, the problem is that national norms may not be the most appropriate reference group for all interpretations of students' scores. When making inferences about aptitude, instead of merely comparing ELL students to all children in the nation who are the same age or in the same grade in school, *compare the scores of ELL students to those of other ELL students* who are the same age or in the same grade. Such additional comparisons take into consideration the ELL student's *opportunity to learn* the skills that were assessed by the test.

When screening classes for academic talent, it is best to test all students for the same aptitudes. Then, to identify those who have the greatest potential in particular domains, compare their performance to that of other students who had roughly similar opportunities to develop the abilities measured. The common practice of administering a nonverbal test to some students and relying on *national norms* misses the majority of these students with special academic talents, especially as with the *Raven* or *NNAT* when the norms tables wrongly assign high scores to many students. The problem is further compounded when additional tests that are administered are normed on different populations and the highest score on all of the tests is inappropriately taken as the best indicator of a student's ability. Even when norms can be trusted, the highest score in a series is generally one of the most error-laden scores (Lohman & Korb, 2006). To reduce many of these problems, the nonverbal test score may be used as part of a more comprehensive identification system. A comprehensive system incorporates a broader range of abilities and teacher ratings, and it formalizes the process of comparing students with their peers rather than comparing them only with the national norm group (Lohman & Renzulli, 2007; Renzulli, 2005).

Implications for Educators

Studies 1 and 2 have several implications for educators. The first is that *it is wrong to assume that nonverbal tests level the playing field for children who come* *from different cultures or who had different educational opportunities*. The much lower performance of ELL students in Study 2 could not be attributed to demographic factors. Nor could it be attributed to an inability to understand the test directions, since the directions were given in both Spanish and English. One plausible explanation is that performance on nonverbal tests depends, in part, on the sophistication of the students' language development. In particular, students use language while thinking about the test items (for example, to label stimuli, to remember rules, and to monitor their work). This affects student performance on a broad range of tasks, especially as those tasks increase in complexity and tax students' working-memory resources.

The second implication is that *educators need to be skeptical about national norms, especially when they administer tests normed on different populations*. The unwary educator who administers the *Raven* or *NNAT* to students who previously took *CogAT* would mistakenly assume that these tests identified many gifted children that *CogAT* missed. For the *Raven*, the primary differences are in the mean scores; for *NNAT*, the primary differences are in the variability of the scores. Since identifying gifted students depends critically on both the mean and the variability of the distribution, *many more students will obtain unusually low or unusually high scores on the Raven and NNAT than the norms tables lead users to expect*.

A related implication is that *those who use tests to identify students with exceptional scores need to look at the distributions of scores on the tests that they use*. This is not difficult to do. *CogAT* users can request these distributions from the test publisher. Also, districts administering *CogAT* and using the *Interactive Results Manager*™ can obtain score distributions with a few mouse clicks. It was only by examining score distributions that we discovered that the most common score on *NNAT* for ELL students in grade 1 was a stanine of 1 (the lowest possible stanine) and that the most common score on the *Raven* for English-proficient students in grades 1 and 2 was a stanine of 9 (the highest possible stanine).

Third, *Study 2 did not support several claims commonly made about differences between NNAT and other nonverbal tests in its ability to identify academically talented minority students*. For example, there was no evidence that *NNAT* identifies equal proportions of high-scoring students from different ethnic or language groups (Naglieri & Ford, 2003). Rather, the differences between Caucasian students and their Black, Hispanic, Asian-American, and American Indian classmates were large both at the mean and in the proportions of high-scoring students. Neither did the study find differences of only one point on *NNAT* between ELL and English-proficient students, even after controlling for ethnicity, SES, and other demographic factors. Rather, the difference was 10 points.

Fourth, in spite of these limitations *a good nonverbal reasoning test can help identify bright children*, especially those who come from low-SES families or who are not fluent in the language of the dominant culture. However, the identification of talent is best made from measures that are closer to the specific cognitive, affective, and conative aptitudes required for success in the available educational programs, rather than from tests that do not measure these critical aptitudes. Students who might someday excel as writers, mathematicians, or artists will obtain high scores on verbal, quantitative, or spatial tests that measure these specific aptitudes. But their development will not be considered unusual unless their test scores are compared to the test scores of other children who have had roughly similar opportunities to develop the abilities being measured. This applies to all abilities – even those measured by nonverbal reasoning tests.

Those of us who have the privilege of working on *CogAT* know that it sets the standard not only for the assessment of verbal reasoning and quantitative reasoning but also for the assessment of nonverbal reasoning. The battery of nonverbal tests that Robert Thorndike and Elizabeth Hagen developed many years ago and then refined and updated through six revisions is unequalled in psychometric quality and educational utility. Most educators are aware of the resources for score interpretation and use that accompany *CogAT*. However, differences in psychometric quality are less apparent. For example, the errors of measurement on the three *CogAT* batteries are approximately half as large as the corresponding errors of measurement on *NNAT* or *OLSAT*.[7] The studies summarized in this newsletter show another important but fundamental difference between the Nonverbal Battery of *CogAT* and other nonverbal tests: the quality and dependability of the test norms.

[7] Indeed, the standard errors of measurement for *CogAT* are smaller than the standard errors of measurement for corresponding scores from individually administered ability tests.

References

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

Cattell, R. B., & Cattell, K. S. (1965). *Manual for the Culture-Fair Intelligence Test, Scale 2*. Champaign, IL: Institute for Personality and Ability Testing.

George, C. E. (2001). *The Naglieri Nonverbal Ability Test: Assessment of cross-cultural validity, reliability, and differential item functioning*. (Doctoral dissertation, Fordham University, 2001). *Dissertation Abstracts International*, 62, 6018.

Laing, P. C., & Castellano, J. (2007). *Overview of Project Bright Horizon*. Paper presented at the Annual Meeting of the National Association of Gifted Children, Minneapolis.

Lohman, D. F. & Korb, K. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted*, 29, 451-484.

Lohman, D. F., Korb, K. A., & Lakin, J. (under review). Identifying academically gifted English Language Learners: A comparison of the *Raven*, *NNAT*, and *CogAT*. *A copy of this paper may be downloaded from* http://faculty.education.uiowa.edu/dlohman/.

Lohman, D. F., & Renzulli, J. (2007). *A simple procedure for combining ability test scores, achievement test scores, and teacher ratings to identify academically talented children.* A copy of this paper may be downloaded from http://faculty.education.uiowa.edu/dlohman/.

Mulligan, J. L. (2007). *Assessment of giftedness: A concise and practical guide*. New York: YBK Publishers.

Naglieri, J. A., Booth, A. L., & Winsler, A. (2004). Comparison of Hispanic children with and without limited English proficiency on the *Naglieri Nonverbal Ability Test. Psychological Assessment*, *16*, 81-84.

Naglieri, J. A., & Ford, D. Y. (2003). Addressing underrepresentation of gifted minority children using the Naglieri Nonverbal Ability Test (NNAT). *Gifted Child Quarterly*, 47, 155-160.

Naglieri, J. A., & Ronning, M. E. (2000). The relationship between general ability using the Naglieri Nonverbal Ability Test (NNAT) and Stanford Achievement Test (SAT) reading achievement. *Journal of Psychoeducational Assessment*, *18*, 230-239.

Renzulli, J. S. (2005). *Equity, excellence, and economy in a system for identifying students in gifted education: A guidebook* (RM05208). Storrs, CT: The National Research Center on the Gifted and Talented.

Shaunessy, E., Karnes, F. A., & Cobb, Y. (2004). Assessing potentially gifted students from lower socioeconomic status with nonverbal measures of intelligence. *Perceptual and Motor Skills*, *98*, 1129-1138.

Tannenbaum, A. (1965) [Review of the Culture Fair Intelligence Test.] In O.K. Buros (Ed.), *The sixth mental measurement yearbook* (pp.721-723). Highland Park, NJ: The Gryphon Press.

Riverside Publishing
A HOUGHTON MIFFLIN COMPANY
www.riversidepublishing.com
800.323.9540

Cognitively Speaking is a publication of Riverside Publishing.

Riverside Publishing
3800 Golf Road, Suite 100
Rolling Meadows, IL 60008

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