Gifted Child Quarterly
Volume 52 Number 4
Fall 2008 275-296

10.1177/0016986208321808
http://gcq.sagepub.com
hosted at
http://online.sagepub.com

Identifying Academically Gifted English-Language Learners Using Nonverbal Tests

A Comparison of the Raven, NNAT, and CogAT

David F. Lohman
Katrina A. Korb
Joni M. Lakin
University of Iowa

Abstract: In this study, the authors compare the validity of three nonverbal tests for the purpose of identifying academically gifted English-language learners (ELLs). Participants were 1,198 elementary children (approximately 40% ELLs). All were administered the Raven Standard Progressive Matrices (Raven), the Naglieri Nonverbal Ability Test (NNAT), and Form 6 of the Cognitive Abilities Test (CogAT). Results show that the U.S. national norms for the Raven substantially overestimate the number of high-scoring children; that because of errors in norming, the NNAT overestimates the number of both high-scoring and low-scoring children; that primary-level ELL children score especially poorly on the NNAT; that the standard error of measurement was twice as large for the NNAT as for the Raven or the CogAT; that ELL children scored .5 to .67 standard deviations lower than non-ELL children on the three nonverbal tests; and that none of the nonverbal tests predict achievement for ELL students very well.

Putting Research to Use: Do nonverbal reasoning tests level the field for ELL children? Many practitioners have assumed that they do. However ELL children in this study scored 8 to 10 points lower than non-ELL children on the three nonverbal tests. The study also shows that practitioners cannot assume that national norms on the tests are of comparable quality. When put on the same scale as CogAT, Raven scores averaged 10 points higher than CogAT and NNAT scores. For NNAT, the mean is correct but the variability was up to 40% too large. Thus, when using national norms, both the Raven and NNAT will substantially overestimate the number of high-scoring children.

Keywords: *identification; English-language learners; nonverbal tests*

English-language learners (ELLs) are underrepresented in programs that serve gifted students. Because of this, program administrators have searched for alternative procedures for identifying academic talent that would increase the representation of these children in their programs (Ford & Harris, 1999; Frasier, García, & Passow, 1995). Most of these procedures rely heavily or exclusively on group-administered, nonverbal reasoning tests.

Nonverbal tasks have long formed an important part of both individual intelligence tests such as the Wechsler Scales (e.g., Wechsler, 1949) and group ability tests such as the Lorge-Thorndike Intelligence Tests (Lorge, Thorndike, & Hagen, 1964). Scores on the nonverbal batteries of these tests provided one indicator of ability for native speakers of the language but often served as the only measure of ability for examinees who were not fluent speakers of the language. The nonverbal test enabled examiners to interpret the scores of all examinees—both those with language difficulties and those without such difficulties—using the same norms tables.

The question, then, is not whether nonverbal tests should be administered to ELL children. All would agree that such tests can provide helpful information. Rather, the issue is whether nonverbal tests should provide

the only estimate of ability or if other measures of ability should be used to provide additional information about a student's academic aptitude. Additional measures are not needed if nonverbal tests can adequately capture the aptitude constructs of interest.

Differential psychologists have long cautioned that nonverbal reasoning tests do not capture the same ability construct that is measured by tests that use language (Anastasi & Urbina, 1997) and therefore should not be used alone to make decisions about academic giftedness (Terman, 1930) or general intellectual competence (J. Raven, Raven, & Court, 1998; McCallum, Bracken, & Wasserman, 2001).

Even though nonverbal reasoning tests may be good measures of general ability (*g*), they do not measure the specific verbal and quantitative abilities that add importantly to the prediction of academic success for students from all ethnic backgrounds (Gustafsson & Balke, 1993; Keith, 1999; Lohman, 2005b). Webb, Lubinski, and Benbow (2007) argue that in addition to verbal and quantitative reasoning, spatial ability should also routinely be assessed in talent searches. However, nonverbal tests such as Raven's Progressive Matrices Test (J. C. Raven, 1941), the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997), and the Nonverbal Battery of the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001) measure fluid reasoning ability, not spatial ability. Good tests of spatial ability require examinees to create and transform visual images, for example by mentally rotating images in three-dimensional space. Such tests show substantial sex differences. However, the Raven, NNAT, or the CogAT do not require these skills and therefore do not show significant sex differences. These non-verbal reasoning tests are not effective for identifying students who excel in visual-spatial thinking (Lohman, 1994).

Nevertheless, those who use nonverbal tests to help identify academically gifted students are generally more interested in identifying students who might excel in traditionally structured academic programs than in measuring visual thinking abilities. But there is a tradeoff. In the language of test score validity, nonverbal reasoning tests can reduce the amount of construct-irrelevant variance in test scores for nonnative speakers by reducing the impact of language. This enhances validity. But not measuring the ability to reason in verbal or quantitative symbol systems, underrepresents the construct of fluid reasoning ability and therefore reduces the validity of the test scores. (Braden, 2000). How practitioners might negotiate this tradeoff is one of the issues we hope to address in this study.

Estimating Score Differences Between ELL and Non-ELL Children

In spite of concerns about construct underrepresentation, nonverbal reasoning tests are sometimes used to screen all students for inclusion in programs for the gifted because it is thought that such tests level the playing field for ELL and non-ELL children. However, studies that compare the performance of ELL and non-ELL children on nonverbal tests are rare, and so it is uncertain how much nonverbal reasoning tests reduce the difference in mean scores of ELL and non-ELL students. Instead, most studies compare the performance of students from different ethnic groups (Lewis, 2001; Stephens, Kiger, Karnes, & Whorton, 1999) rather than ELL and non-ELL children within those ethnic groups. But even these studies have given widely varying estimates of the magnitude of score differences between children from different ethnic backgrounds.

Several reports on the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1996) have examined this issue. For example, in an analysis of the NNAT fall standardization data, Naglieri and Ronning (2000a) reported a 3-point difference on the Nonverbal Ability Index (NAI) scale ($M = 100$, $SD = 15$) between the average scores of Hispanic and White children. However, this difference was obtained after students were equated on socioeconomic status (SES), region of the country, gender, urbanicity (urban or rural area of residence), and school type (public or private school attendance). Controlling for these variables also reduced differences between

Authors' Note: Katrina A. Korb is now at the University of Jos, Nigeria. We gratefully acknowledge the support and assistance of the Project Bright Horizon Research Team, particularly Peter Laing who supervised the monumental task of collecting the data for this project. The views and opinions expressed in this article are those of the authors and should not be ascribed to any members of the Project Bright Horizon staff or its consulting partners. The research reported in this paper was sponsored by a Jacob K. Javits Gifted and Talented Education Grant Program to the Project Bright Horizon Research Team: Peter Laing, Project Director/Co–Principal Investigator, Washington Elementary School District, Phoenix, AZ; Dr. Jaime Castellano, Project Consultant; and Dr. Ray Buss, Arizona State University at the West Campus, Principal Investigator. Please address correspondence to David F. Lohman, Belin-Blank Center, 600 Honors Center, University of Iowa, Iowa City, IA 52242; e-mail: david-lohman@uiowa.edu.

Note: This article was accepted under the editorship of Paula Olszewski-Kubilius.

Hispanic and White children on the Mathematics Battery of the Stanford Achievement Test (1995) by the same amount. This suggests that equating students on demographic variables equated them on other factors as well. Furthermore, the practice of statistically controlling for the effects of social class and other environmental variables on a presumably cultural fair test is inherently illogical. If the test were in fact culture fair, then such controls would be unnecessary.

Although controlling unwanted sources of variation can clarify relationships between variables, it can also obscure them. For example, SES is commonly defined by three variables: family income, parental education, and parental occupational status. Therefore, controlling for SES also controls for that portion of the ability variance that predicts how much education parents obtain or that may be required for their occupations. However, controlling for parent ability also controls in part for the abilities of the parent's biological children. Failure to keep track of the shared, construct-relevant variance when controlling ability test scores for SES is one example of a larger problem called the *partialing fallacy* (see Lubinski, 2000). In addition to the potential for introducing conceptual confusions, the practice of statistically controlling for variables can make it difficult for users to estimate the magnitude of group differences that they might expect to see in their schools.

In a second analysis of the same data—this time not controlling other variables—Naglieri and Ford (2003) reported that the NNAT identified equal proportions of high-scoring White, Black, and Hispanic children as gifted. However, other investigators have not found that the NNAT identified equal proportions of high-scoring students from different ethnic groups, either with groups of Black and White students (Shaunessy, Karnes, & Cobb, 2004; Stephens et al., 1999) or with groups of Hispanic and White students (Drake, 2006; Lewis, 2001). Indeed, an independent analysis of the NNAT standardization data found large differences between the scores of White, Black, and Hispanic students at all ages (George, 2001). On the NAI scale (M = 100, SD = 15), the differences between White and Hispanic students ranged from 9 points at level A to 3 points at Level G, with a median across levels of 6 points. The median Black–White difference was 12 NAI points. The inconsistency between the Naglieri and Ford (2003) report of proportional representation of White, Hispanic, and Black students on the NNAT and that of other investigators who have used the same or similar tests with these populations has never been explained, despite questions about the integrity of the analyses that were performed on the data (Lohman, 2005a, 2006). Nevertheless, Naglieri (2007) asserts that the Naglieri and Ford (2003) paper is one of the most important studies on the NNAT.

In a third analysis of the NNAT standardization data that once again controlled for urbanicity, SES, region of the country, type of school, and gender, Naglieri, Booth, and Winsler (2004) reported a 1-point difference on the NAI scale between ELL and non-ELL Hispanic children. This comparison directly addresses how much the scores of ELL and non-ELL children might differ—at least within the world in which students do not differ on these demographic variables. Unfortunately, there was no external criterion for identifying ELL students in the many schools that participated in the test standardization. Even when given explicit criteria for identifying ELL students, schools differed widely on how they interpreted the criteria (Lutkus & Mazzeo, 2003). Therefore, some ELL children may have been included in the non-ELL group and vice versa, thereby underestimating the difference between ELL and non-ELL students. Furthermore, it is unclear how large the difference between ELL and non-ELL students might be if one or more of the five demographic variables were not controlled.

There is also uncertainty about the magnitude of differences between ELL and non-ELL Hispanic students on the Standard Progressive Matrices (Raven; J. C. Raven, Court, & Raven, 1996). Although at least one study found no differences between the performance of Hispanic and White students on the Raven (e.g., Powers, Barkan, & Jones, 1986), most investigators report differences of .5 to .7 *SD* (Hoffman, 1983; Mills & Tissot, 1995; Saccuzzo & Johnson, 1995). Of course, many Hispanic children are not ELLs, and so the Hispanic–White difference mostly likely underestimates the size of the difference between Hispanic ELL children and White non-ELL children, especially when these groups also differ in socioeconomic status (J. Raven, 1989). A larger problem for selection decisions, however, is that the Raven has never been properly normed on the U.S. population. This has led to considerable confusion about the interpretability of normative scores on the test.

For Form 6 of the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001), there have been even fewer reports of ethnic differences. However, differences between ethnic groups were estimated for analyses of item bias (Lohman & Hagen, 2002). Within each school in the standardization sample, all students belonging to a particular minority group

were identified. Then an equal number of majority students were randomly sampled from the same school. Differences in the mean scores for these two groups of students were averaged across all schools in the national sample. One advantage of this method is that it provides an unbiased estimate of the size of the test score difference that test users are likely to see within their schools. Differences between White and Hispanic students averaged 5 Standard Age Score (SAS) points on the Nonverbal Battery (M = 100; SD = 16). When adjusted for differences in score scale (SD = 16 on CogAT versus SD = 15 on NNAT), these differences between Hispanic and White students are approximately the same as those obtained by Naglieri and Ronning (2000a) on the NNAT. We know of no other reports of score differences between White and Hispanic students on the CogAT.

A major difficulty with all of these studies is that the category *Hispanic* includes students from diverse cultural backgrounds with markedly different English-language skills. For example, Lohman (2005b) reported that Hispanic students with at least two high scores (stanines of 8 or 9) on the three CogAT batteries (Verbal, Quantitative, and Nonverbal) were no more likely than Black students, and even less likely than Asian American students, to show a profile of lower verbal reasoning abilities. These high-ability Hispanic students performed more like other ethnic minority students than like ELL students. This reinforces the need to separate the influences of ethnicity and ELL status on observed score differences.

Finally, just because group differences are larger on one test (e.g., the CogAT Verbal Battery) than on another test (the CogAT Nonverbal Battery) does not mean that the latter test can better identify the most academically talented students. The problem is that means and correlations can tell quite different stories about the data.

Means Versus Correlations

Two traditions have dominated the long history of quantitative research in education and psychology (Cronbach, 1957). Each of these methods considers as error the variance that the other method hopes to explain. On the one side are those who study differences between the mean scores of groups—usually groups formed by random assignment of individuals to treatments. Differences among individuals within treatment groups are considered random error. On the other side are those who study correlations among scores on different measures. The individual differences that are considered error when means are compared represent the systematic variation that the correlational method hopes to explain.

When applied to the same data, these two perspectives—analyses of means versus analyses of correlations—can lead to different interpretations. For example, although individuals vary widely in height, males are on average taller than females. However, the genetic factor that explains the difference between the average heights of males and females (i.e., the presence of a Y chromosome) cannot explain individual differences in height within either group.

There is a similar paradox here: Although manipulations that reduce the impact of construct-irrelevant sources of difficulty can enhance construct validity, the same manipulations can also reduce construct validity by curtailing the extent to which the test measures important aspects of cognition that define the construct. For example, by reducing the language demands of tests, one can reduce the difference between the mean scores of ELL and non-ELL students. However, nonverbal tests measure a narrower range of cognitive abilities and thus show lower correlations with measures of academic accomplishment than do ability tests that also assess students' reasoning abilities in those symbol systems (verbal and quantitative) most needed for success in academic learning (Lohman, 2005b; Mills & Tissot, 1995). Importantly, when properly computed within ethnic groups, the magnitude of these correlations does not differ across ethnic groups. Correlations between nonverbal, figural reasoning abilities and reading achievement typically range from r = .4 to .5; correlations with mathematics achievement typically range from r = .5 to .6 (Lohman & Hagen, 2002; Naglieri & Ronning, 2000b; Powers et al., 1986). Although significant, these correlations are considerably smaller than the correlation of r = .8 between verbal reasoning and reading achievement or between quantitative reasoning and mathematics achievement (Lohman & Hagen, 2002; Thorndike & Hagen, 1995). Because lower predictive validity can substantially impair the ability of the test to identify academically talented students, college admissions tests such as the SAT have continued to measure verbal and quantitative reasoning abilities, even though questions about test bias would be easier to address with a nonverbal test.

Recognizing these limitations, several investigators have counseled caution in the use of nonverbal tests for screening children for inclusion in programs for the gifted (e.g., Bittker, 1991; Lohman, 2005b; Mills

& Tissot, 1995). They note that nonverbal tests violate the dictum that abilities assessed by the identification procedures should correspond with those that are required for success in the program (Mills & Tissot, 1995; Renzulli, 2005). Academic learning requires verbal, quantitative, and spatial abilities in addition to *g*, and so by this argument, tests that measure academic aptitude must measure all of these abilities.

Purposes of the Study

Clearly then there is controversy both about how best to identify academically talented minority students and about the efficacy of different nonverbal tests for achieving this goal. The primary goal of this study was to compare the performance of a large sample of ELL and non-ELL children on three of the most widely used nonverbal tests: the Standard Progressive Matrices Test, the NNAT, and the Nonverbal Battery of Form 6 of the CogAT. In addition to analyses of group differences, the relative effectiveness of the three nonverbal tests in identifying those ELL and non-ELL children who displayed the strongest academic achievement was also examined. Academic achievement, of course, is not synonymous with either the broad construct of giftedness or with the narrower construct of academic giftedness. Nevertheless, good achievement tests provide a useful indicator of academic talent.

Several controls were implemented to enhance the validity of the study. These included counterbalancing the order of administration of the three tests, using only trained examiners, testing children in the familiar surroundings of their regular classrooms, giving directions in Spanish or English (as appropriate), and securing the collaboration of the authors of two of these tests (Dr. Naglieri and Dr. Lohman) in the design of the study. Each collaborating partner was then given a copy of the data to analyze.

Specifically, we asked the following questions: (a) Are the normative scores similar for each test? (b) How large are the differences between ELL and non-ELL children on each test? (c) Do the tests identify similar proportions of high-scoring ELL and non-ELL students or of students from different ethnic groups? (d) Are the tests sufficiently reliable to make decisions about giftedness? and (e) When compared to measures of verbal and quantitative reasoning, how well do the nonverbal tests identify the most academically successful ELL students?

Method

Participants

In all, 2,087 students were administered at least one nonverbal test. Only the 1,198 that completed all three nonverbal tests were included in these analyses. Some characteristics of these students are reported in Table 1. All attended grades K to 6 in two elementary schools in a large Southwestern school district in the United States. Students were classified as New English-language learners (NELL) or Continuing English-language learners (CELL) based on the type of services they were receiving and their scores on the Stanford English Language Proficiency Test (SELP; Harcourt Educational Measurement, 2003). Except at kindergarten (where all ELL students were classified as NELL), 80 to 88% of the ELL students at each grade were classified as CELL. The native language of almost all ELL students was Spanish. Almost all (95.4%) of the 786 Hispanic children, 90.8% of the 164 students from other minority groups, and 53.2% of the 248 White students were eligible for free or reduced lunch.

Measures

Raven's Standard Progressive Matrices. J. C. Raven devised the Progressive Matrices Test to measure the eductive component of Spearman's g. The essential feature of eductive ability is "the ability to generate new, largely nonverbal, concepts" (J. C. Raven et al., 1996, p. 1). The most recent version of the Standard Progressive Matrices uses the same items as the 1938 version of the test (J. C. Raven, 1941). This test consists of five sets of 12 problems that follow a common theme. Each item requires students to examine the components of an incomplete matrix and then to select the empty box that best completes the matrix.

Administration of the Raven required approximately 60 minutes. Because the Raven has only one level, students at all grades took the same test. Students recorded their answers on a separate answer sheet, which was then scored by hand.

Naglieri Nonverbal Ability Test. The NNAT is described as "a brief, culture-fair, nonverbal measure of school ability" (Naglieri, 1997, p. 1). The test uses a figural matrix format similar to the Raven. There are several differences between the NNAT and the Raven. First, items on the NNAT have five rather than six or

Table 1
Number of Students, by Gender, Ethnicity, and ELL Status

	Gender		Ethnicity			ELL	
Grade	Male	Female	White	Hispanic	Other	New	Continuing
K	63	62	27	82	16	65	0
1	130	100	48	148	34	23	91
2	100	90	44	113	33	15	59
3	81	90	26	121	24	12	69
4	101	90	37	131	23	9	50
5	85	73	36	103	19	5	33
6	62	71	30	88	15	4	30

Note: For ethnicity, *other* consists of 69 Black, 61 American Indian, and 34 Asian American students. ELL = English-language learner.

eight response options. Items are printed in blue, white, and yellow and are clustered into four groups (Pattern Completion, Reasoning by Analogy, Serial Reasoning, and Spatial Visualization). The proportion of items in each cluster varies across levels of the test. Pattern Completion items require examinees to identify the missing portion of a patterned rectangle. Reasoning by Analogy and Serial Reasoning items require examinees to determine how a figure changes across the rows and columns of a design. Spatial Visualization items require that examinees determine how two or more designs combine to create a new figure.

The test is organized into seven levels, each of which contains 38 different items. The recommended NNAT level was administered for each grade. Students marked their answers directly in test booklets for Levels A through C and on a separate, machine-readable answer sheet at Levels D and E.

Cognitive Abilities Test (Form 6). Form 6 of CogAT consists of three separate batteries that measure verbal, quantitative, and nonverbal reasoning (Lohman & Hagen, 2001). Although each battery can be administered alone, all three batteries were administered in this study. The Primary Edition of CogAT (Levels K, 1, and 2) is designed for students in kindergarten through second grade. Each of the three primary batteries has 40, 44, and 48 items at Levels K, 1, and 2, respectively. The items in each battery are divided into two subtests with different item formats. No reading is required. Children listen to the teacher read a question and then choose the picture that best answers the question. For the Verbal Battery, the subtests are Oral Vocabulary and Verbal Reasoning; for the Quantitative Battery, they are Relational Concepts and Quantitative Concepts; and for the Nonverbal Battery, they are Matrices and Figure Classification. The Matrices subtest follows the same general format as the items on the Raven and the NNAT. The Figure Classification subtest presents three figures in the stem. The student must select the fourth figure that belongs to the set.

The Multilevel Edition of CogAT6 (Levels A to H) is typically administered to students in Grades 3 through 12. The Multilevel Verbal (65 items), Quantitative (60 items), and Nonverbal (65 items) batteries each contain three subtests that use different item formats. The student must read individual words on two subtests of the Verbal Battery (Verbal Analogies and Verbal Classification) and a sentence on the third (Sentence Completion). The three subtests of the Quantitative Battery are Number Series, Quantitative Relations, and Equation Building. The three subtests of the Nonverbal Battery are Figure Classification, Figure Analogies, and Figure Analysis. The Figure Classification subtest presents three figures in the stem, and the examinee is required to determine a fourth figure that belongs to the set. Figure Analogies contains three figures in an analogy (A→B : C→__) that the student must complete. Figure Analysis requires the examinee to determine how a folded, hole-punched paper will appear when unfolded.

The recommended level of CogAT was administered at each grade. Students marked their answers directly in test booklets for Levels K, 1, and 2 and on a separate, machine-readable answer sheet for Levels A to D.

Achievement Tests. The Arizona Instrument to Measure Standards Dual Purpose Assessment (AIMS DPA) was designed to yield normative and criterion-referenced information about student achievement. Thirty to 50% of the items on the AIMS DPA were taken from the TerraNova achievement tests (CTB/McGraw-Hill, 2002). The remaining items were developed by educators specifically for the AIMS DPA to better align the test with state educational goals (Arizona Department of Education, 2006). A combined Reading/Language Arts subtest and the Mathematics subtest of the AIMS DPA each contained approximately 80 items. Separate scores are reported for Reading, Language, and Mathematics. In this study, national grade percentile ranks based on the norm-referenced TerraNova items were reported as part of the description of the student sample (see Table 2). Composite reading and mathematics scores

Table 2
Median National Grade Percentile Ranks on Form 6 of the Cognitive Abilities Test (CogAT) Verbal, Quantitative, and Nonverbal Batteries; the Naglieri Nonverbal Abilities Test (NNAT); the Standard Progressive Matrices Test (Raven); and the TerraNova Mathematics, Reading, and Language Tests, by ELL Status and Grade

		CogAT					TerraNova		
Grade	*n*	V	Q	N	NNAT	Raven[a]	Math	Reading	Lang
					Non-ELL				
K	44-60	14	20	46	38	83			
1	113-116	32	40	57	55	85			
2	114-116	38	44	64	53	82			
3	82-90	32	43	40	50	75	39	36	40
4	121-132	39	33	46	51	75	43	41	44
5	115-120	34	38	55	50	70	39	48	51
6	94-99	37	44	58	59	70	49	52	43
					ELL				
K	55-65	8	11	27	27	62			
1	107-114	6	14	42	35	63			
2	73-74	7	12	35	37	62			
3	76-81	10	20	24	39	60	20	14	15
4	58-59	6	10	25	34	53	18	18	18
5	37-38	5	14	18	35	41	17	12	14
6	33-34	4	12	17	44	44	11	14	20

Note: ELL = English-language learner; V = Verbal; Q = Quantitative; N = Nonverbal; Lang = Language.
[a] National Age Percentile Rank. Grade percentile ranks not available.

that combined items from both assessments (the full AIMS DPA) were used when investigating the relationships between scores on the three ability tests and student achievement. (Refer to Table 7 in the Results section.)

Stanford English Language Proficiency Test. The Stanford English Language Proficiency (SELP; Harcourt Educational Assessment, 2003) test is based on the standards developed by the Teachers of English to Speakers of Other Languages and measures the English listening, reading, comprehension, writing, and speaking skills of K-12 ELL students. The SELP tests are untimed and group administered except for the speaking portion of the test, which is administered individually. Administration time is typically less than 2 hr.

Procedure

The three ability tests were administered by trained examiners to intact classes in counterbalanced order in late April and early May of 2006. Directions for the tests were given in Spanish or English as appropriate. Each of the three nonverbal tests was administered in a single session separated by approximately 1 week. The Verbal and Quantitative batteries of CogAT were also administered to all children, but in separate sessions from the Nonverbal Battery. These sessions were generally conducted during the same week in which the CogAT Nonverbal Battery was administered. Analyses were performed using SPSS 14.0 (SPSS Inc., 2005).

Results

Judgments about exceptionality depend importantly on the quality and recency of the test norms, the normality of the score distributions, and the reliability and validity of the test scores. Therefore, we first report basic descriptive statistics, score distributions, reliabilities, and the proportions of ELL versus non-ELL students at each score stanine for each of the three tests. Ability tests that aim to identify the most academically talented students should identify many of the students who currently excel academically. Therefore, in the second part of the Results section, we examine correlations between the nonverbal assessments with each other and with measures of reading and mathematics achievement. These address the basic issue of predictive validity for the ability tests.

Table 3
Means (Standard Deviations) for All Students on Form 6 of the Cognitive Abilities Test (CogAT) Nonverbal Battery, the Naglieri Nonverbal Abilities Test (NNAT), and the StandardProgressive Matrices Test (Raven), by ELL Status and Grade

	CogAT Nonverbal SAS		NNAT NAI		Raven Ability Index		
Grade	M	SD	M	SD	M	SD	*n*
			Non-ELL				
K	96.4	(13.0)	94.8	(19.2)	112.4	(13.7)	60[a]
1	102.9	(15.9)	101.6	(20.9)	116.6	(16.1)	116[a]
2	106.3	(12.8)	104.1	(16.9)	116.4	(14.3)	116
3	96.6	(15.5)	99.6	(16.7)	110.4	(15.8)	90
4	97.8	(13.0)	100.3	(15.5)	110.1	(15.6)	132
5	101.6	(13.8)	98.1	(12.9)	108.2	(14.2)	120
6	100.4	(14.4)	103.8	(14.0)	107.9	(15.4)	99
All grades	100.7	(14.4)	100.7	(16.8)	111.5	(15.5)	733
			ELL				
K	91.6	(15.5)	88.5	(18.2)	103.9	(11.5)	65[a]
1	98.0	(14.5)	91.3	(20.2)	106.6	(17.1)	114[a]
2	95.1	(13.7)	90.1	(17.2)	105.2	(16.1)	74
3	89.9	(13.8)	93.3	(15.2)	103.9	(16.7)	81
4	88.9	(11.1)	90.4	(17.1)	101.2	(13.6)	59
5	85.9	(10.0)	88.9	(12.0)	99.0	(16.6)	38
6	86.4	(12.3)	91.1	(14.2)	98.6	(14.4)	34
All grades	92.4	(14.1)	90.7	(17.2)	103.4	(15.8)	465
All Students	97.5	(14.9)	96.8	(17.6)	108.5	(16.1)	1198

Note: For the CogAT and the Raven, population *SD* = 16; for the NNAT, population *SD* = 15. SAS = Standard Age Score.
[a] Because students aged 6 and younger were excluded from Raven norms, *n* = 17 for non-ELL in Grade K; 91 for non-ELL, Grade 1; 24 for ELL, Grade K; and 89 for ELL, Grade 1.

Descriptive Statistics

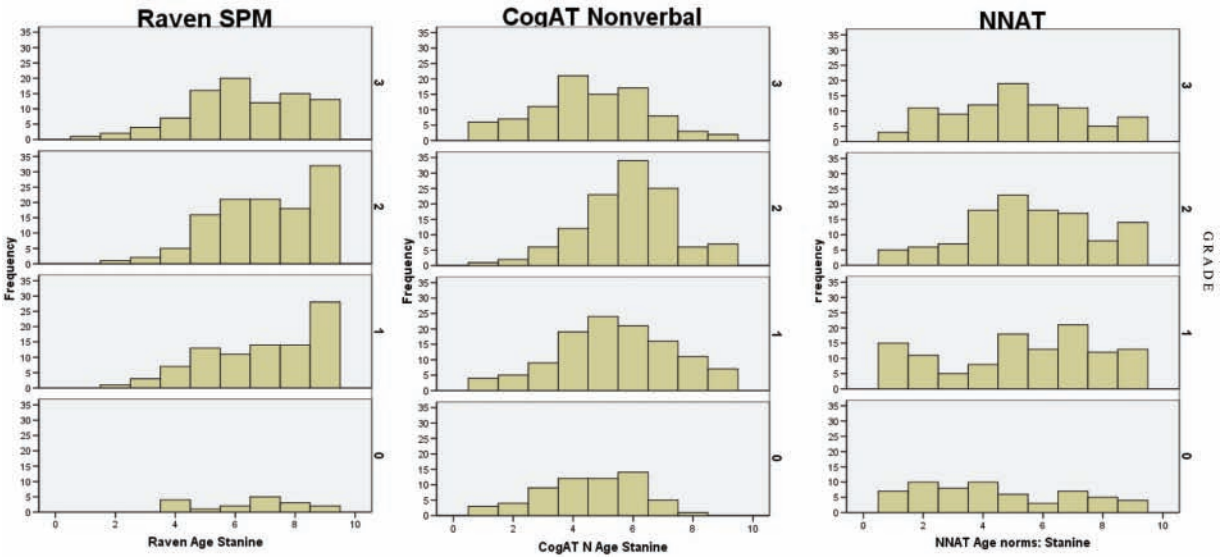
Comparisons with achievement tests. Comparisons between the three nonverbal tests and the TerraNova (CTB/McGraw-Hill, 2000) reported in Table 2 provide both information on the characteristics of the sample and a comparison of norms on the tests. Grade percentile ranks (PRs) were used for all tests except the Raven, which reports only age percentile ranks. However, median age and grade percentile ranks for the CogAT and NNAT were similar, as is the case when students' ages are typical for their grades. Non-ELL children performed at or somewhat below the national average and ELL students considerably below the national average. For example, the median PRs on the Mathematics Battery ranged from 39 to 49 for non-ELL and from 11 to 20 for ELL children. Percentile ranks on the CogAT Quantitative Battery were generally similar.

Percentile ranks on the Raven were considerably higher than the CogAT Nonverbal or the NNAT scores except for ELL children in grades 5 and 6. For non-ELL children, Raven percentile ranks were 25 points higher than both CogAT and NNAT percentile ranks. Although it is possible that the percentile ranks for the NNAT and the CogAT Nonverbal are both too low, this is unlikely, given the congruence between the CogAT Quantitative Battery and the Mathematics subtest of the TerraNova. Rather, the Raven norms appear to be far too lenient.

Nonverbal tests only: All ELL and non-ELL students. Means, standard deviations, and sample sizes for each of the three nonverbal tests are reported separately for ELL and non-ELL students by grade in Table 3. For the CogAT, Standard Age Scores (M = 100, SD = 16) are reported; for the NNAT, Nonverbal Ability Index (NAI) scores (M = 100, SD = 15) are reported. For the Raven, a comparable score dubbed the Raven Ability Index (RAI; M = 100, SD = 16) was constructed using the national age percentile ranks from the 1986 U.S. norms (Table RS3SPM6 in J. C. Raven, 1990).[1]

The mean score for ELL students was substantially lower than the mean score for non-ELL students on all three tests. The means for the CogAT Nonverbal and the NNAT were similar (ELL: *M* = 92 for CogAT and *M* = 91 for NNAT; non-ELL: *M* = 101 for both CogAT and NNAT). However, scores on the Raven were about 11 points higher than the other two tests (*M* = 103 and *M* = 112 for ELL and non-ELL students, respectively).

Figure 1
Non–English-Language Learners



Frequency distributions of national age stanines at grades K (0), 1, 2, and 3 for non–English-language learners on the Standard Progressive Matrices (Raven; J. C. Raven et al., 1996), the Nonverbal Battery of Form 6 of the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001), and the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1996). See Table 1 for sample sizes.

When the data were examined by grade, the pattern of mean scores was similar for the three tests for non-ELL students but different for ELL students. Raven scores were substantially higher than either CogAT or NNAT scores at all grades. The next largest difference was between NNAT scores and CogAT Nonverbal scores for Grade 1 and Grade 2 ELL students: NNAT scores for these students were significantly lower than CogAT Nonverbal scores by 6.7 points at Grade 1, t (113) = 5.19; $p < .001$, and 5.8 points at Grade 2, t (73) = 3.91; $p < .001$. However, there were no differences between scores on these two tests for non-ELL children at Grades 1 and 2. Finally, NNAT scores were somewhat higher than CogAT scores at Grades 3 to 6, although the differences were smaller and only sometimes statistically significant.

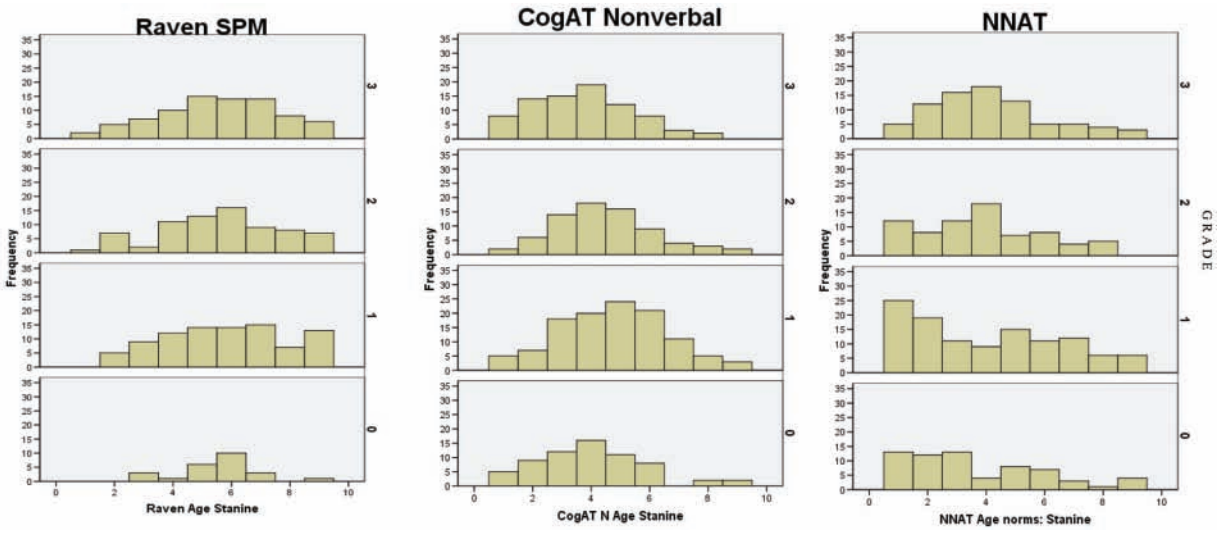
Differences in variability. By design, the population SD is 16 for the CogAT Nonverbal SAS and the Raven RAI score and is 15 for the NNAT NAI score. Because it is unlikely that the two schools that participated in this study represent the full range of ability in the U.S. population, we would expect the SDs in this sample to be somewhat smaller than their population values. This was the case for the CogAT and the Raven. Overall SDs for ELL and non-ELL children were both below the population value of 16. For the NNAT, on the other hand, the overall SD of 17.2 for ELL children was significantly greater than the population value of 15; χ^2 (464) = 611, p < .0001. The SD of 16.8 for non-ELL children was also significantly greater than the population value of 15, $\chi^2(732) = 915$, p < .0001.

When examined by grade, none of the *SD*s for either the CogAT Nonverbal or the Raven were significantly larger than the population value of 16. For the NNAT, however, *SD*s both for ELL and non-ELL children in grades K, 1, and 2 were all significantly greater than the population *SD* of 15.

Score distributions. Giftedness is an inference about ability that is made when scores fall in the upper tail of a score distribution. Differences in the variability of score distributions can therefore change inferences about exceptionality. The number of students who score above a particular value depends both on the shape of the score distribution and on the norms that are used. If the norms do not represent the population, then the number of students who exceed the cutoff may be unexpectedly high (or low).

To better understand the unexpectedly large *SD*s for the NNAT and its impact on giftedness classifications, we constructed histograms of age stanines (based on national norms) for all three tests, separately by

Figure 2
English-Language Learners



Frequency distributions of national age stanines at grades K (0), 1, 2, and 3 for English-language learners on the Standard Progressive Matrices (Raven; J. C. Raven et al., 1996), the Nonverbal Battery of Form 6 of the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001), and the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1996). See Table 1 for sample sizes.

grade and ELL status. We created histograms for students in grades K to 3, because these grades showed the greatest variability in both *M*s and *SD*s across the three tests (see Table 3). Distributions of age stanines are shown in Figure 1 for non-ELL students and in Figure 2 for ELL students. Sample sizes for the Raven are somewhat smaller at grades K and 1 because children aged 6 and younger are excluded from the norms tables. Although RAI scores could not be computed for these children, their data were included in all analyses that used raw scores.

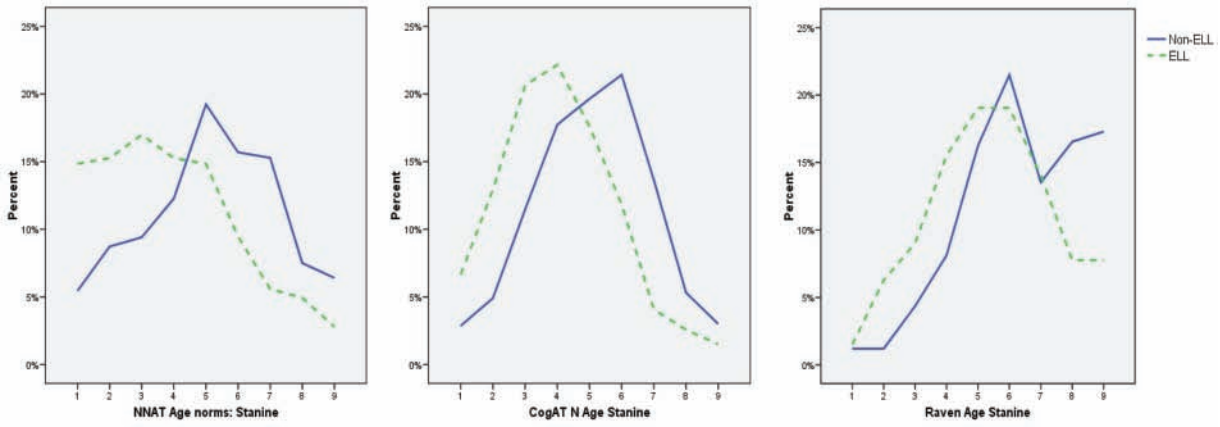
The histograms in Figures 1 and 2 show that CogAT scores were approximately normally distributed for both ELL and non-ELL students at all grades. This was not the case for either the Raven or the NNAT. Extremely high stanine scores were much more common on the Raven than on either the NNAT or the CogAT. Indeed, the modal stanine was 9 on the Raven for non-ELL children in Grades 1 and 2. This indicates that the 1986 U.S. Raven norms result in far too many students being identified as gifted. However, for ELL students in grades K, 1, and 2 and for non-ELL students in grades K and 1 taking the NNAT, the opposite pattern was evident: For these students, extremely low scores were more common than expected. In fact, the modal stanine on the NNAT was only 1 for ELL students in Grade 1.

The preponderance of low scores in the NNAT distributions for ELL children in Figure 2 may indicate that some children did not understand what they were supposed to do, despite the fact that test directions were given in Spanish when appropriate. However, this does not explain the excess of high-scoring children in the non-ELL distributions for grades 1 and 2 (see Figure 1). Instead, the generally flatter and more variable NNAT score distributions could reflect a problem in the initial scaling or norming of the test or simply could reflect more error of measurement in the scores. Whatever its cause, the increased variability of the NNAT results in many more students being classified as very high or very low ability than one would expect in a normally distributed population with an *SD* of 15.

Proportion of ELL and Non-ELL students at each stanine. Another way to compare tests in terms of their impact on the identification of ELL students for gifted programs is in terms of the proportion of ELL versus non-ELL students at different points in the score distribution. The percentage of children at each stanine for the three tests are shown in Figure 3, separately for ELL and non-ELL children. Comparison of the two lines in each plot shows the extent to which the test identifies similar proportions of ELL and

Figure 3

Percentage of students at each stanine on the Naglieri Nonverbal Ability Test (NNAT; left panel), the Cognitive Abilities Test Nonverbal Battery (CogAT N; center panel), and the Standard Progressive Matrices (Raven; right panel) for ELL students (dashed line) and non-ELL students (solid line).



non-ELL children. If the proportions are the same, then the two lines would be coincident. For NNAT (leftmost panel), the critical feature is the preponderance of ELL students (dashed line) with very low scores. Raven stanine scores (rightmost panel) show the opposite pattern: a preponderance of non-ELL students with high scores. The CogAT stanine scores (center panel) show two essentially normal distributions, one to the left of the other. However, for all three tests, proportionately fewer ELL children obtained higher stanine scores.

ELL versus Non-ELL Hispanic students. One of the key questions that we sought to address in this study was the magnitude of score differences between ELL and non-ELL Hispanic children on these three tests. Naglieri et al. (2004) reported a difference of only 1 scale point on the NNAT between Hispanic children with and without limited English proficiency. This difference was obtained after samples were first equated on SES (approximated by whether the child was on free or reduced lunch and by the average educational level of parents in the school district), region of the country, gender, urbanicity (urban or rural area of residence), and school type (public or private school attendance). By design, all of these variables were controlled in this study as well: Children in this study all attended public schools, lived in the same city, and of necessity, in the same region of the country. In addition, more than 95% of the Hispanic children were eligible for free or reduced lunch. Other studies have shown that scores for elementary school children do not vary by gender on any of the three tests (Lohman & Hagen, 2002; J. Raven et al., 1998; Rojahn & Naglieri, 2006).

Table 4 shows the *M*s (and *SD*s) on each of the three tests for ELL and non-ELL Hispanic children, by grade. Although there was some variation across grades, differences between ELL and non-ELL Hispanic children were large on all three tests. Across grades, the average differences were 7.5, 6.3, and 9.5 points on the Raven, CogAT, and NNAT or effect sizes of .47, .46, and .63, respectively. This is a much larger disparity between ELL and non-ELL Hispanic students than the 1-point difference reported by Naglieri et al. (2004). Even though the ELL and non-ELL Hispanic children were similar in many respects, the ELL Hispanic children at all grades were on average less able to cope with the demands of the tests than were non-ELL Hispanic children.

Typically, large differences at the mean translate into much larger discrepancies in the odds of obtaining extreme scores (Feingold, 1994; Hedges & Nowell, 1995). However, the large differences between the mean scores for ELL and non-ELL Hispanic children were offset by the somewhat greater variability of scores for ELL students, especially on the NNAT. Once again, this illustrates the importance of understanding how distributions of scores differ across tests for particular groups of examines.

Table 4

Means and Standard Deviations for 786 Hispanic Children on the Form 6 Nonverbal Battery of the Cognitive Abilities Test (CogAT), the Naglieri Nonverbal Ability Test (NNAT), and the Standard Progressive Matrices Test (Raven), by Grade and ELL Status

	CogAT Nonverbal SAS			NNAT NAI Score[a]			Raven Ability Index Score[b]		
Grade	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
K									
Non-ELL	99	12.1	21	99	17.9	21	114	13.3	8
ELL	93	14.6	61	90	17.8	61	105	10.8	23
1									
Non-ELL	100	16.7	43	99	18.9	43	112	14.6	34
ELL	97	14.3	105	91	20.5	105	106	17.5	82
2									
Non-ELL	102	10.3	46	100	14.5	46	111	14.3	46
ELL	96	12.5	67	91	16.8	67	106	15.4	67
3									
Non-ELL	98	15.6	45	102	16.2	45	111	16.0	45
ELL	90	13.8	76	93	15.2	76	104	16.5	76
4									
Non-ELL	98	11.6	74	101	14.6	74	110	15.3	74
ELL	89	11.3	57	91	17.1	57	101	13.1	74
5-6									
Non-ELL	100	12.2	125	101	12.3	125	108	13.5	125
ELL	86	10.1	66	90	11.5	66	99	15.1	66
Total									
Non-ELL	100	12.9	354	100	14.8	354	110	14.5	332
ELL	92	13.7	432	91	17.1	432	104	15.7	371

Note: ELL = English-language learner; NAI = Nonverbal Ability Index; SAS = Standard Age Score.

[a] Population $M = 100$, $SD = 15$.

[b] The Raven Ability Index (RAI) score was derived from the 1986 U.S. National norms reported in the test manual (J. C. Raven, 1990). The score is normally distributed with $M = 100$ and $SD = 16$ to correspond with the CogAT SAS.

Ethnic differences. The final group comparison examined the proportion of students in each ethnic group who received a stanine score of 9 on one or more of the three nonverbal tests. To control for differences that might be due to language familiarity rather than ethnicity, only the scores of non-ELL students were used in this analysis. If Naglieri and Ford (2003) are correct, then the NNAT should show approximately equal proportions of children from each ethnic group in the ninth stanine. The relevant data are shown in Figure 4. Equal proportions of high-scoring students from each ethnic group would appear as a horizontal line in the figure. Clearly, this was not observed for any of the tests. Although the pattern of scores was similar, the height of the profile varied as a function of the total number of students across all ethnic categories whose scores were assigned a stanine of 9 by the national norms tables for the test. Asian American and White students were much more likely to obtain stanine scores of 9 than were American Indian, Hispanic, or Black students on all three tests. For the Raven, 33.3% of the Asian students and 28.0% of the White students received stanine scores of 9. The corresponding percentages for the NNAT were 15.0 and 12.4; for the CogAT Nonverbal, 5.0 and 7.0. At the other extreme, the percentage of Black students was 3.7, 1.6, and 0 on the Raven, NNAT, and CogAT Nonverbal, respectively.

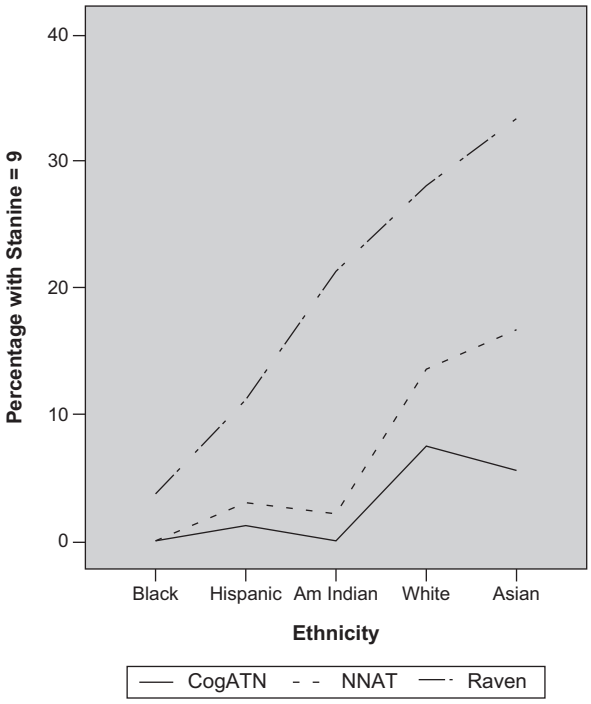
Individual Differences

Analyses to this point have focused on comparisons of score distributions, particularly on their means and variances. In this section, we examine the dependability of individual differences on the three tests and the consistency of these individual differences across the three ability tests and relationships with two measures of academic achievement.

Reliabilities. An essential property of all assessments is the reliability of the scores derived from them. But reliability is a function not only of the test and the procedures that are used to estimate the coefficient but

Figure 4

Percentage of Non–English-language learners who obtained a national age stanine score of 9 on each of the three nonverbal tests, by ethnicity.



also of the characteristics of the sample of examinees. Although they have little impact on group means, differences in score reliability can substantially affect score distributions. Therefore, the differences we observed between ELL and non-ELL children in average performance or the unexpectedly wide variation in scores on the NNAT might in part reflect differences in the reliabilities of the tests—overall or for ELL children in particular.

To test this hypothesis, we estimated the standard error of measurement (SEM) on the NNAT and on the CogAT for students in this sample. Item scores were not available for the Raven (which had to be hand scored), and so it could not be included in the analysis.[2] However, raw score scales differ for the CogAT and the NNAT, and so we transformed these raw score SEMs to scale score SEMs. We used the same *SD* for both the NNAT and the CogAT, so that the estimated standard errors would be on a common scale.

Analyses were first performed separately for ELL and non-ELL students. However, the results were the same for both groups. Therefore the differences in previous analyses between ELL and non-ELL cannot be attributed to differences in the consistency of their behavior across items within each test. Instead, differences between tests—especially between the NNAT and the Raven or the CogAT Multilevel Edition—might in part reflect differences in reliability. In fact, one of the more important influences on reliability is the number of items. The NNAT has 38 items at all levels. The same 60 Raven items are presented at all grades. The CogAT Nonverbal has 40, 44, and 48 items at levels K, 1, and 2, respectively, and 65 items at all other levels. Because it has almost twice as many items as the NNAT at grades 3 and higher, the CogAT Nonverbal should be more reliable than the NNAT at those levels.

Reliability coefficients for raw scores and the corresponding SEM of the scale scores for the CogAT and the NNAT are reported in Table 5, separately for each grade. Reliability coefficients can be misleading when test score variances differ. For example, the NNAT reliability coefficient at Grade 2 ($r_{xx'}$ = .89) is larger than the reliability coefficient at Grade 3 ($r_{xx'}$ = .84). However, the SEMs show the opposite pattern (8.2 at Grade 2 vs. 6.6 at Grade 3) because the variance of test scores is larger at Grade 2 than at Grade 3. This is one reason that measurement experts advocate using the SEM on the reported score scale (here, NAI or SAS units) rather than the reliability coefficient when scores on a test are interpreted (Anastasi & Urbina, 1997; Feldt & Brennan, 1989).[3]

Across grades, the SEM for the NNAT was typically more than twice as large as the SEM for CogAT. The table also shows how SEMs influence the width of a 68% confidence interval for an SAS of 129. This corresponds to a percentile rank of 97, a common criterion for decisions about academic giftedness. For the NNAT, the median 68% confidence interval across grades was 14 points; for the CogAT Nonverbal, 6 points. The actual confidence intervals for both tests would surely be even larger for high-scoring students. Error of measurement commonly doubles or triples as scale scores near the extremes of the distribution (Feldt & Brennan, 1989). Therefore, one possible contributor to the nonnormal score distributions on the NNAT is greater error of measurement in the scores, particularly for those students with extremely high or low scores.

Correlations among the ability tests. Correlations bring all scores to a common scale and discard information about differences in means or variances. Therefore, even though normative scores (e.g., percentile ranks, SAS, NAI, and RAI scores) on the three

Table 5
Standard Deviations, KR 20 Reliability Coefficients, Standard Errors of Measurement (SEM), and 68% Confidence Intervals for a Standard Age Score (SAS) of 129 on the Nonverbal Battery of Form 6 of the Cognitive Abilities Test (CogAT) and a Nonverbal Ability Index (NAI) Score of 129 on the Naglieri Nonverbal Ability Test (NNAT), by Grade

	CogAT Nonverbal SAS				NNAT NAI Scores			
Grade	*SD*	KR 20	SEM	68% CI for 129	*SD*	KR 20	SEM	68% CI for 129
K	15.32	.97	3.1	126-132	19.18	.91	5.6	123-135
1	15.04	.95	3.5	126-133	21.93	.92	6.6	122-136
2	14.08	.93	3.7	125-133	21.26	.89	8.2	121-137
3	15.72	.97	3.1	126-132	17.79	.84	6.6	122-136
4	13.58	.96	3.0	126-132	17.90	.88	7.7	121-137
5	14.59	.96	3.2	126-132	13.15	.80	5.3	124-134
6	15.16	.96	3.2	126-132	14.71	.85	5.6	123-135

Note: The 68% confidence interval for an SAS of 129, a commonly used cutoff score for admittance to gifted programs, was used. For the NNAT, SEMs in this table assume that the population *SD* of NAI scores is 16 rather than 15. This was done so that the SEMs for the CogAT and the NNAT would be directly comparable. To convert these SEMs to their true values, multiply each by 15/16.

nonverbal tests are not interchangeable, it is possible that individual differences on these normative scores, when brought to a common scale, show considerable congruence. Table 6 reports correlations among raw scores on the three ability tests, separately for ELL and non-ELL students. Correlations were pooled across grades. Although the correlations were slightly higher for non-ELL than for ELL students, all fell within the range of $r = .60$ to .65. This is about the level that would be expected for figural reasoning tests that measure similar constructs. For example, correlations among the three figural reasoning tests on the Multilevel Edition of the CogAT Nonverbal Battery are generally in the $r = .6$ to .7 range (Lohman & Hagen, 2002).

As might be expected, correlations between the three nonverbal scores and the CogAT Verbal Battery were lower for ELL students than for non-ELL students. Correlations with the CogAT Quantitative Battery were intermediate. In fact, the correlation between the CogAT Nonverbal and Quantitative batteries was only slightly smaller ($r = .60$) for ELL students than for non-ELL students ($r = .69$). This supports the recommendation that admissions committees consider scores on both the CogAT Verbal and the CogAT Quantitative-Nonverbal (QN) partial composite when they are identifying academically talented students (Lohman & Renzulli, 2007).

Consistency of identification. Those who must rely on tests to identify gifted students commonly underestimate the degree of inconsistency that will be observed even when tests are highly correlated (Lohman & Korb, 2006). Given three tests that measure a similar construct and that correlate $r = .6$ to $r = .65$, what percentage of the students would be identified as gifted by at least two tests? Consider the 1,064 students who had age stanine scores on all three nonverbal tests. Of this group, stanine scores of 9 were obtained by 146 students on the Raven, 51 on the NNAT, and 26 on the CogAT Nonverbal. One would expect most of the much smaller group of students who received a stanine of 9 on the NNAT and all of the even smaller group who obtained a stanine of 9 on the CogAT to be included in the much larger group of 146 who obtained a stanine of 9 on the Raven. However, only 36 from the NNAT group and 18 from the CogAT group did so. And only 11 received a stanine of 9 on all three tests. There is no gold standard, which is why—when students are to be identified for admission to special programs—scores on several assessments that measure the same construct should be put on the same scale and then averaged. Admitting those students with a high normative score on any one of the tests increases the errors of measurement and regression to the mean (Lohman & Korb, 2006).

Verbal and Quantitative abilities. Although the aim of this study was to compare the three nonverbal tests, we requested that the CogAT Verbal and Quantitative batteries also be administered. As expected, differences between ELL and non-ELL students were much smaller on the three nonverbal tests (Raven, NNAT, and CogAT-Nonverbal) than on the Verbal and Quantitative batteries of CogAT. Across grades, the

Table 6
Pooled Within-Grade Correlations Among Raw Scores for Non–English-Language Learners and English-Language Learners on Form 6 of the Cognitive Abilities Test (CogAT), the Naglieri Nonverbal Ability Test (NNAT), and the Standard Progressive Matrices Test (Raven)

	CogAT			NNAT	Raven
Test	Verbal	Quantitative	Nonverbal		
		Non-ELL (n = 664)			
CogAT					
Verbal	1.00	0.71	0.64	0.44	0.50
Quantitative		1.00	0.69	0.55	0.56
Nonverbal			1.00	0.65	0.62
NNAT				1.00	0.66
Raven					1.00
		ELL (n = 426)			
CogAT					
Verbal	1.00	0.53	0.40	0.31	0.35
Quantitative		1.00	0.60	0.48	0.48
Nonverbal			1.00	0.60	0.61
NNAT				1.00	0.64
Raven					1.00

Note: ELL = English-language learners.

average difference between ELL and non-ELL students was 16.6 SAS points on the CogAT Verbal Battery and 13.2 points on the Quantitative Battery. Thus, the differences between ELL and non-ELL students were twice as large on the CogAT Verbal Battery (16.6 points) as on the CogAT Nonverbal Battery (8.3 points). But does this mean that the nonverbal test is a better measure of academic talent?

Predictive validity. Correlations between the three ability tests and measures of reading and mathematics achievement are shown in Table 7. Correlations with the reading and mathematics achievement scores are reported separately by grade and ELL status. The achievement tests were only administered at grades 3 through 6. The Grade 5 and 6 samples were combined in an effort to obtain a sufficiently large sample of ELL students.[4]

There were several noteworthy results. First, correlations were uniformly higher for non-ELL students than for ELL students. There are several ways to interpret this finding. For example, it could mean that the ability tests are less valid for ELL students, that the achievement tests are less valid (or reliable) for ELL students, or that some ELL students responded much more to instruction than other ELL students between the time the ability and achievement tests were administered.

Second, none of the nonverbal tests predicted reading achievement very well. Except for the Grade 5-6 ELL sample, the CogAT Verbal Battery was a much better predictor of reading achievement for both ELL and non-ELL children. The median correlations for the three nonverbal tests with reading achievement were r = .49 and .35 for non-ELL and ELL children, respectively. Interestingly, the median correlation between NNAT and reading comprehension in Spanish was also r = .35 in a separate study (Naglieri & Ronning, 2000a). Thus, that the TerraNova reading test was in English does not appear to be the cause of the lower correlation between the nonverbal tests and reading comprehension for ELL students.

The correlations between the CogAT Verbal and reading achievement were considerably higher: r = .76 and .54 for non-ELL and ELL students, respectively. This represents a substantial increase in predictive validity. However, as shown by the VQN multiple correlations in Table 7, the CogAT Nonverbal scores added little or nothing to the prediction of reading comprehension afforded by the CogAT Verbal scores alone (again, with the exception of the Grade 5-6 ELL group). Indeed, the regression weight for the Nonverbal score hovered around zero—sometimes positive, sometimes negative—for both ELL and non-ELL students.

Third, although the three nonverbal tests better predicted mathematics achievement than reading achievement, the CogAT Quantitative showed higher correlations. However, as shown in Table 7, the best prediction was obtained when CogAT Verbal and Nonverbal were also entered into the regression. This is a common finding. Learning mathematics and performing well on mathematics achievement tests require verbal reasoning and figural–spatial reasoning as well as quantitative reasoning ability (Floyd, Evans, & McGrew, 2003). This applies to all students—those who are native speakers of English and those who are learning to speak the language.

Discussion

This controlled comparison of the Raven, the NNAT, and the CogAT showed that the three tests differ importantly in the quality of their norms, in the reliability of the scores they produce, and in their ability to identify the most academically able ELL and non-ELL students.

First, and most important, we observed substantial differences between the nonverbal test scores of ELL

Table 7

Correlations Between Form 6 of the Cognitive Abilities Test (CogAT), the Naglieri Nonverbal Ability Test (NNAT), and the Standard Progressive Matrices Test (Raven), the Stanford English Language Proficiency Test (SELP) and Composite TerraNova Reading and Mathematics Achievement Test Scores, by English-Language Learner Status and Grade

	Non-ELL			ELL		
Test Score	Grade 3	Grade 4	Grade 5-6	Grade 3	Grade 4	Grade 5-6
	Reading Composite Scale Score					
CogAT						
Verbal SS	.76	.79	.74	.68	.54	.27
Quantitative SS	.62	.68	.61	.43	.48	.30
Nonverbal SS	.56	.61	.52	.35	.38	.29
VQN Multiple R	.77	.81	.76	.68	.59	.36
NNAT Scale Score	.35	.49	.49	.38	.35	.16
Raven Raw Score	.42	.55	.47	.46	.39	.19
SELP Oral Prof				.45	.49	.18
	Mathematics Composite Scale Score					
CogAT						
Verbal SS	.74	.75	.67	.61	.54	.54
Quantitative SS	.78	.82	.78	.68	.62	.62
Nonverbal SS	.73	.75	.68	.56	.44	.44
VQN Multiple R	.84	.87	.82	.73	.67	.58
NNAT Scale Score	.55	.60	.66	.48	.38	.38
Raven Raw Score	.55	.60	.58	.46	.47	.47
SELP Oral Prof				.51	.50	.50
n	82-90	120-131	205-215	74-79	53-58	69-70

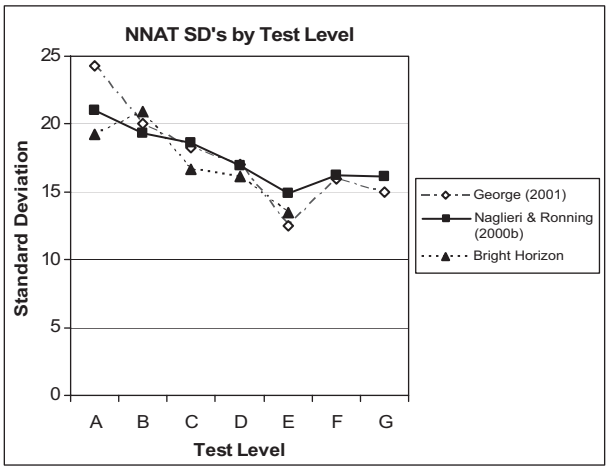
Note: ELL = English-language learners; SS = Scale Score.

and non-ELL children, both on average and in the proportions of these children at different points in the score distributions. These proportions varied dramatically across the three tests. On the NNAT, ELL students were much more likely to receive very low scores. On the Raven, non-ELL children were much more likely to receive very high scores. Only the CogAT Nonverbal showed normally distributed scores for both groups. Differences between the mean scores of ELL and non-ELL children were reduced only slightly when we controlled for ethnicity by comparing only Hispanic ELL and non-ELL students. Previous research that statistically controlled for environmental factors found small differences between ELL and non-ELL students on the NNAT (Naglieri et al., 2004). However, this study naturally controlled for the same variables and found differences of 7.5, 7.3, and 9.5 points on the Raven, CogAT, and NNAT, respectively. These differences are congruent with the conclusion that nonverbal tests do not see through the veneer of culture, education, or language development (Anastasi & Urbina, 1997; Cronbach, 1990; Scarr, 1994).

Second, norms for two of the nonverbal tests were seriously awry. In particular, the 1986 U.S. norms for the Raven were, on average, markedly easier than the 1995-1996 NNAT norms or the 2000 CogAT norms. When placed on the same scale as CogAT, scores for both ELL and non-ELL students were 10 to 11 points higher on the Raven than on either the NNAT or the CogAT Nonverbal tests. Higher scores on Raven in part reflect the increase in scores on all ability tests (especially figural reasoning tests) that has occurred over the past 70 years (Dickens & Flynn, 2001; Flynn, 1999; J. Raven, 2000). But in larger measure, it most likely reflects the fact that the 1986 U.S. norms for the Raven were not based on a representative national sample but on a compilation of convenience samples of test scores for the 20 to 26 schools or school districts that submitted scores to the test authors over the years.

For the NNAT, on the other hand, NAI scores were much more variable than they should have been, especially at the primary grades. Distributions of NAI scores showed that very low scores on the NNAT were much more common than expected for ELL

Figure 5
Standard deviations for *NNAT* from (1) George (2001), (2) Naglieri and Ronning (2000b), and (3) the Project Bright Horizon Study.



students in grades K to 2 and for non-ELL students at Grade 1. On the other hand, high NAI scores were more common than expected for non-ELL students. We explored the possibility that the excessively large *SD*s might reflect the fact that the SEM for the NNAT was typically more than twice as large as the SEM for the Raven or the CogAT Nonverbal. However, a consistently larger SEM cannot explain why the variance of NAI scores increased systematically as one moved from level E downward to level A.

Therefore, we looked for other published reports that might show the same broad dispersion of NNAT NAI scores. George (2001) reanalyzed the Spring NNAT standardization data. She reported *SD*s for number correct scores at each level. We used these raw score *SD*s to estimate *SD*s of NAI scores. Naglieri and Ronning (2000b) reported *SD*s of NAI scores using the Fall NNAT standardization data.

These two sets of *SD*s for the NNAT standardization data are plotted in Figure 5 along with the SDs from the Project Bright Horizon study. All three data sets show the same pattern of decreasing *SD*s across test levels. If NAI scores had been properly standardized, then all of these SDs would be approximately 15.

Inferences about giftedness depend critically on the *SD* of scores. Excessive variability of NAI scores means than the test will overidentify the number of students receiving high scores. The extent to which the NNAT overidentifies the number of high-scoring

Table 8
Overidentification Rates for the Number of Students With Nonverbal Ability Index (NAI) Scores Above 115, 130, and 145

	True NAI Score		
Test 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

students is shown in Table 8. 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 more than 3 times as many gifted children as the properly normed test.

Third, in our analyses of test score validity, we examined the extent to which the different tests correlated with each other and were able to predict reading and mathematics achievement. Although normative scores on the three nonverbal tests were not interchangeable, all three appeared to measure a common ability dimension. In terms of predictive validity, we found that correlations for all tests were higher for non-ELL than for ELL students and that the best predictors of achievement were given by the students' abilities to reason in the symbol systems most essential for learning in that domain. For reading comprehension, this was the CogAT Verbal score, whereas for mathematics, it was a weighted combination of the three CogAT scores.

Why does the CogAT Verbal score predict reading achievement? Could it reflect common reading demands? Beginning at Grade 3, students must read individual words on the Verbal Classification and Verbal Analogies subtests on the Verbal Battery and a short sentence on each item in the Sentence Completion subtest. However, these reading demands are minimal when compared to the demands of the reading comprehension test. The typical passage on the reading subtest of the TerraNova has approximately 300 words at Grades 3 to 6. This means that students must read about 1,800 words across the six passages. This excludes reading the questions, options, and task directions. Furthermore, by design,

the reading level of most words on the CogAT Verbal Battery is well below grade level, and the estimated readability of the sentences is unrelated to their difficulty.[5] Perhaps the best explanation for the correlation between measures of verbal reasoning and reading comprehension is one of the oldest: Reading comprehension is an exercise in verbal reasoning (Thorndike, 1917). Furthermore, the variegated, ill-structured concepts that can be represented by words differ qualitatively from the well-structured concepts represented on figural reasoning tests. Nonverbal reasoning tests do not capture this kind of reasoning. Indeed, if anything, the unique aspects of verbal and figural or spatial reasoning interfere with one another (Lohman, 1994).

As in studies that compared different ethnic groups (Keith, 1999; Lohman, 2005b), the pattern of the correlations between ability and achievement tests did not differ for ELL and non-ELL students. This means that the identification of academic talent requires measurement of the same aptitude variables for all children. What it does not mean is that all children should be compared to a common norm group, especially if the goal is to identify talent rather than to label giftedness. Rather, inferences about talent (or aptitude) require the simple step of comparing children's performance to that of other children who have had roughly similarly opportunities to develop the abilities measured by the test. Separating test scores for all ELL and all non-ELL children is no more difficult than separating the scores for boys and girls or for third-graders and fourth-graders.

Recall that SAS scores for ELL students on the CogAT Nonverbal Battery were much higher than their SAS scores on the other two CogAT test batteries. But SAS scores compare children's test scores to all other children in the nation. For these ELL students who were just learning the English language, these SAS scores mean that their performance was approximately 1 year behind their non-ELL classmates on the Quantitative Battery and approximately 2 years behind on the Verbal Battery. Although this is useful information, it is unhelpful for making inferences about aptitude or about the rate at which a child is acquiring competence compared to others with similar levels of experience. One gets a very different picture when the same test scores are compared to those of other ELL children. Now half of the ELL children have normative scores above the mean! And some have very high normative scores.

In conclusion, large differences between the scores of ELL and non-ELL children on the three nonverbal tests show that one must consider opportunity to learn not only for tests that measure verbal and quantitative abilities and achievements but also for those abilities measured by nonverbal tests. The common practice of administering a nonverbal test only to a fraction of the population and then relying on national test norms has no doubt masked these differences, especially when the norm tables wrongly assign high scores to many students. The problem is further compounded when—in an effort to identify more students—additional tests that were normed on different populations are administered and the highest score on any test is inappropriately taken as the best indicator of the student's ability (see Lohman & Korb, 2006). Many of these problems can be attenuated by using the nonverbal test score as one part of a comprehensive identification system that incorporates a broader range of abilities and teacher ratings and that formalizes the process of comparing students with their peers rather than with a distant and often inadequate national norm group (Lohman & Lakin, 2007; Lohman & Renzulli, 2007; Renzulli, 2005).

Limitations

This study has several limitations. First, although the sample was large and diverse in ethnicity, grade, and ELL status, it was not representative of all Hispanic ELL students (much less, all ELL students) in the United States. Second, the criterion information was limited to achievement test scores. Performance on achievement tests is a narrower concept than *academic giftedness*, which in turn is narrower than *giftedness*. A wider range of both predictor and outcome measures would have been helpful in the identification of students whose talents are not well captured by achievement tests used in the schools. Although some teacher ratings were gathered, rating scales were not consistent across schools and thus could not be used in these analyses. Nonverbal reasoning tests might better predict performances in domains that also require reasoning about visual stimuli such as design or model building, although even here good measures of spatial ability would probably show greater validity (Gohm, Humphreys, & Yao, 1998; Smith, 1964).

A longitudinal study would provide stronger evidence on the utility of different tests for identifying children who excel academically at some point in the future or who profit differentially from different kinds of educational interventions such as acceleration or enrichment. However, this limitation may be remedied if state assessment results are gathered (as

planned) for those students in this study who remain in the state in future years.

Implications

This study has several important implications for educators. First, one cannot assume that nonverbal tests level the playing field for children who come from different cultures or who have had different educational opportunities. The ELL children in this study scored from .5 to .6 *SD* lower than non-ELL children on all three nonverbal tests. The lower performance of ELL children could not be attributed to economic factors, to the student's age or grade, or to other demographic factors. Nor could it be attributed to an inability to understand the test directions, because directions were given in Spanish whenever necessary.

The second implication is more of a caution. Practitioners need to be appropriately skeptical about national norms, especially for tests that were normed on different populations. The unwary user who administered the Progressive Matrices or the NNAT after administering a test with good norms would incorrectly assume that these tests were much more successful in identifying gifted children than the first test. Because of outdated or improperly computed normative scores, many more students will obtain unusually high (and, on the NNAT, unusually low) scores than the norms tables would lead one to expect.

A related implication is that practitioners should always examine the distributions of test scores. This is not difficult to do. Many teachers construct histograms by hand for scores on classroom tests. If asked, most test publishers will report these distributions and local norms derived from them. However, anyone who can use a basic spreadsheet (such as Microsoft Excel) can do it with a few mouse clicks on the data sets that publishers provide. It was only by examining score distributions that we discovered that the most common score for ELL children in Grade 1 was a stanine score of 1 on the NNAT or that the most common score for non-ELL children in both grades 1 and 2 was a stanine score of 9 on the Progressive Matrices Test.

Finally, this controlled comparison of the Raven, NNAT, and CogAT provided no support for the claim that the NNAT identifies equal proportions of high-scoring students from different ethnic or language groups. Nonverbal tests need not fulfill a utopian vision as measures of innate ability unencumbered by culture, education, or experience in order to play a useful role in the identification of academically gifted children. Nonverbal reasoning tests do help identify bright children, especially those who are poor or who are not fluent in the language of the dominant culture. When combined with measures of quantitative reasoning and spatial ability, nonverbal reasoning tests are particularly effective for identifying students who will excel in engineering, mathematics, and related fields (Shea, Lubinski, & Benbow, 2001). As this example illustrates, the identification of talent in any domain is best made from measures that are more proximal to the specific cognitive, affective, and conative aptitudes required for success in that domain than from measures of more distal constructs. Students who might someday excel as writers, mathematicians, or artists will generally show rapid learning when given the opportunity to learn concepts and skills in those domains. These students will also obtain high scores on the verbal, quantitative, or spatial tests that measure the specific aptitudes required to develop competence in the domain (Corno et al., 2002). 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 that are measured (Lohman & Lakin, 2007). This applies to all abilities—even those abilities measured by nonverbal reasoning tests.

Notes

1. This was accomplished by finding the score in a normal distribution with a mean of 100 and *SD* of 16 that corresponded with each percentile rank. For example, percentile ranks of 16, 50, and 84 correspond with RAI scores of 84, 100, and 116, respectively.

2. For each test, we first estimated KR 20 reliability ($r_{xx'}$) from the item scores and then computed the error variance V(e) by $V(e) = V(x)(1 - r_{xx'})$, where $V(x)$ is the variance of the raw scores. For the CogAT, this was done separately for each subtest and the error variances summed (Feldt & Brennan, 1998). The square root of this total error variance gives the raw score SEM. Our best guess is that the SEM for Raven would be approximately 2.7 to 3.0 on a scale with $M = 100$ and $SD = 16$. Oddly, SEMs are only reported graphically in the Raven test manual.

3. The reliability coefficient is estimated by the ratio of two variances:

$$\text{Reliability Coefficient} = \frac{\text{Observed Score Variance} - \text{Error Variance}}{\text{Observed Score Variance}}$$

A large increase in observed score variance is generally associated with a much smaller increase in error variance. Indeed, the error variance is often assumed to be constant across samples.

4. Even though scale scores are used for all tests except the Raven (for which raw scores on the same test form are comparable across grades), combining the grade 5 and 6 samples in this way may have introduced other factors, especially for the achievement tests. This could explain the lower than expected correlations for the Grade 5-6 group.

5. Students who have reading difficulties can be administered the test orally. The assertion that the CogAT Verbal Battery is just a reading test cannot explain why the reliability of differences between the CogAT Verbal score and ITBS Reading Comprehension is about $r = .7$. If the two tests measured the same thing, then this reliability would be 0. Nor is it true that items are difficult to read. Naglieri and Ford (2005) calculated the readability of items on Level D (i.e., Grade 6) of the CogAT Sentence Completion subtest using the Flesch-Kincaid Grade Level method (Flesch, 1948). However, readability formulas require passages with a minimum of 100 words. Applying these formulas to sentences produces a result that is mostly random noise. Thus, it is unsurprising that these readability numbers have no relationship with the actual difficulties of the items (Lohman, 2005b). What is surprising is that the readability numbers continue to be presented as dependable facts (Naglieri, 2007), even after these basic misuses of the readability statistics were pointed out. As the late Senator Moynihan once observed, "Everyone is entitled to his or her own opinion, but not to his or her own facts."

References

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

Arizona Department of Education. (2006, November). *AIMS student guides*. Retrieved September 4, 2007, from http://www.ade.state.az.us/standards/AIMS/AIMSSTGuides/

Bittker, C. M. (1991). Patterns of academic achievement in students who qualified for a gifted program on the basis of nonverbal tests. *Roeper Review, 14*, 65-68.

Braden, J. P. (2000). Perspectives on the nonverbal assessment of intelligence. *Journal of Psychoeducational Assessment, 18*, 204-210.

Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., et al. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow.* Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

CTB/McGraw-Hill. (2002). *TerraNova, the Second Edition.* Monterey, CA: Author.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review, 108*, 346-369.

Drake, K. S. (2006, June). *Gifted services identification report.* Saint Paul, MN: Saint Paul Public Schools, Department of Research, Evaluation and Assessment, Office of Research and Development.

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex Roles: A Journal of Research, 30*, 81-92.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221-233.

Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) Cognitive Abilities and mathematics achievement across the school-age years. *Psychology in the Schools, 40*, 155-171.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54*, 5-20.

Frasier, M. M., García, J. H., & Passow, A. H. (1995). *A review of assessment issues in gifted education and their implications for identifying gifted minority students* (RM95204). Storrs, CT: The National Research Center on the Gifted and Talented. Retrieved September 4, 2007, from the University of Connecticut, Neag Center for Gifted Education and Talent Development Web site: http://www.gifted.uconn.edu/nrcgt/reports/rm95204/rm95204.pdf

Ford, D. Y., & Harris, J. J., III. (1999). *Multicultural gifted education* (Education and Psychology of the Gifted Series). New York: Teachers College Press.

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.

Gohm, C. L., Humphreys, L. G., & Yao, G. (1998). Underachievement among spatially gifted students. *American Educational Research Journal, 35*, 515-531.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*, 407-434.

Harcourt Educational Assessment. (2003). *Stanford English Language Proficiency Test.* San Antonio, TX: Author.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*, 41-45.

Hoffman, H. V. (1983). *Regression analysis of test bias in the Raven's Progressive Matrices for Whites and Mexican-Americans* (Doctoral dissertation, The University of Arizona, 1983). *Dissertation Abstracts International, 44*, 3509.

Keith, T. Z. (1999). Effects of general and specific abilities on student achievement: Similarities and differences across ethnic groups. *School Psychology Quarterly, 14*, 239-262.

Lewis, J. D. (2001). Language isn't needed: Nonverbal assessments and gifted learners. *Proceedings of the Growing Partnership for Rural Special Education Conference.* San Diego, CA. (ERIC Document Reproduction Service No. ED 453026)

Lohman, D. F. (1994). Spatially gifted, verbally, inconvenienced. In N. Colangelo, S. G. Assouline, & D. L. Ambroson (Eds.), *Talent development: Vol. 2. Proceedings from the 1993 Henry B. and Jocelyn Wallace National Research Symposium on Talent Development* (pp. 251-264). Dayton: Ohio Psychology Press.

Lohman, D. F. (2005a). Review of Naglieri and Ford (2003): Does the Naglieri Nonverbal Ability Test identify equal proportions of high-scoring White, Black, and Hispanic students? *Gifted Child Quarterly, 49*, 19-28.

Lohman, D. F. (2005b). The role of nonverbal ability tests in identifying academically gifted students: An aptitude perspective. *Gifted Child Quarterly, 49*, 111-138.

Lohman, D. F. (2006). *Identifying academically gifted children in a linguistically and culturally diverse society*. Invited presentation at the Eighth Biennial Henry B. & Jocelyn Wallace National Research Symposium on Talent Development, University of Iowa, Iowa City. Retrieved August 28, 2007, from The University of Iowa Web site: http://faculty.education.uiowa.edu/dlohman/

Lohman, D. F., & Hagen, E. P. (2001). *Cognitive Abilities Test (Form 6).* Itasca, IL: Riverside.
Lohman, D. F., & Hagen, E. P. (2002). *Cognitive Abilities Test (Form 6): Research handbook.* Itasca, IL: Riverside.
Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in *ITBS* and *CogAT* scores during elementary school. *Journal for the Education of the Gifted*, *29*, 451-484.
Lohman, D. F., & Lakin, J. (2007). Nonverbal test scores as one component of an identification system: Integrating ability, achievement, and teacher ratings. In J. VanTassel-Baska (Ed.), *Alternative assessments for identifying gifted and talented students* (pp. 41-66). Austin, TX: Prufrock Press.
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.* Unpublished manuscript, The University of Iowa. Retrieved August 28, 2007, from The University of Iowa Web site: http://faculty.education.uiowa.edu/dlohman/
Lorge, I., Thorndike, R. L., & Hagen, E. (1964). *The Lorge-Thorndike Intelligence Tests*. Boston, MA: Houghton Mifflin.
Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "Sinking shafts at a few critical points." *Annual Review of Psychology, 51, 405-444.*
Lutkus, A. D., & Mazzeo, J. (2003). *Including special-needs students in the NAEP 1998 Reading Assessment: Part I. Comparison of overall results with and without accommodations* (Report NCES 2003-467). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Retrieved September 4, 2007, from http://www.nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003467
McCallum, R. S., Bracken, B. A., & Wasserman, J. D. (2001). *Essentials of nonverbal assessment.* Hoboken, NJ: Wiley.
Mills, C. J., & Tissot, S. L. (1995). Identifying academic potential in students from under-represented populations: Is using the Raven Progressive Matrices a good idea? *Gifted Child Quarterly, 39*, 209-217.
Naglieri, J. A. (1996). *Naglieri Nonverbal Ability Test.* San Antonio, TX: Harcourt Brace Educational Measurement.
Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test: Multilevel technical manual*. San Antonio, TX: Harcourt Brace Educational Measurement.
Naglieri, J. A. (2007). Traditional IQ: 100 years of misconception and its relationship to minority representation in gifted programs. In J. VanTassel-Baska (Ed.), *Alternative assessments for identifying gifted and talented students* (pp. 67-88). Waco, TX: Prufrock Press.
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. (2000a). Comparison of White, African American, Hispanic, and Asian children on the Naglieri Nonverbal Ability Test. *Psychological Assessment, 12*, 328-334.
Naglieri, J. A., & Ronning, M. E. (2000b). The relationship between general ability using the Naglieri Nonverbal Ability Test (NNAT) and the Stanford Achievement Test (SAT) reading achievement. *Journal of Psychoeducational Assessment, 18*, 230-239.
Powers, S., Barkan, J. H., & Jones, P. B. (1986). Reliability of the Standard Progressive Matrices Test for Hispanic and White-American children. *Perceptual and Motor Skills, 62*, 348-350.
Raven, J. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement, 26*, 1-16.
Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.
Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 1. General overview.* Oxford, UK: Oxford Psychologists Press.
Raven, J. C. (1941). Standardisation of Progressive Matrices, 1938. *British Journal of Medical Psychology, 19*, 137-150.
Raven, J. C. (1990). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Research supplement No. 3* (2nd ed.). Oxford, UK: Oxford Psychologists Press.
Raven, J. C., Court, J. H., & Raven, J. (1996). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 3. Standard Progressive Matrices.* Oxford, UK: Oxford Psychologists Press.
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.
Rojahn, J., & Naglieri, J. A. (2006). Developmental gender differences on the Naglieri Nonverbal Ability Test in a national normed sample of 5-17 year olds. *Intelligence, 34*, 253-260.
Saccuzzo, D. P., & Johnson, N. E. (1995). Traditional psychometric tests and proportionate representation: An intervention and program evaluation study. *Psychological Assessment, 7*, 183-194.
Scarr, S. (1994). Culture-fair and culture-free tests. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 322-328). New York: Macmillan.
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.
Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-Year longitudinal study. *Journal of Educational Psychology, 93*, 604-614.
Smith, I. M. (1964). *Spatial ability: Its educational and social significance*. San Diego, CA: Robert R. Knapp.
SPSS, Inc. (2005). *SPSS 14.0 for Windows* [Computer software]. Chicago: Author.
Stanford Achievement Test (9th ed.). (1995). San Antonio, TX: The Psychological Corporation.
Stephens, K., Kiger, L., Karnes, F. A., & Whorton, J. E. (1999). Use of nonverbal measures of intelligence in identification of culturally diverse gifted students in rural areas. *Perceptual and Motor Skills, 88*, 793-796.
Terman, L. M. (1930). Autobiography of Lewis M. Terman. In C. Murchison (Ed.), *History of psychology in autobiography* (Vol. 2, pp. 297-331). Worcester, MA: Clark University Press.
Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology, 8*, 323-332.

Thorndike, R. L., & Hagen, E. (1995). *Cognitive Abilities Test (Form 5) research handbook*. Chicago: Riverside.
Webb, R. M., Lubinski, D., & Benbow, C. P. (2007). Spatial ability: A neglected dimension in talent searches for intellectually precocious youth. *Journal of Educational Psychology, 99*, 397-420.
Wechsler, D. (1949). *Wechsler Intelligence Scale for Children.* New York: The Psychological Corporation.

David F. Lohman is Professor of Educational Psychology at the University of Iowa. He currently directs the Institute for Research and Policy on Acceleration at the Belin-Blank International Center for Gifted Education and Talent Development. His research interests include the effectiveness of different curricular adaptations for academically talented students, conceptualization and measurement of reasoning abilities, and general issues in the identification and development of talent. Since 1998, he has worked with Elizabeth Hagen on the Cognitive Abilities Test.

Katrina A. Korb is a lecturer in Educational Psychology at the University of Jos in Jos, Nigeria. Her primary research interest is in the development of symbolic numerical thinking.

Joni M. Lakin is a doctoral student in Educational Psychology at the University of Iowa. Her research interests include cognitive and noncognitive predictors of academic achievement and the use of assessments to identify gifted students, especially gifted students who are English-language learners.