

Identifying Academically Talented Minority Students

David F. Lohman
The University of Iowa
Iowa City, Iowa

ABSTRACT

The cultural and socioeconomic diversity of the U.S. school population is now and long has been underrepresented in programs for academically advanced students (see, e.g., Donovan & Cross, 2002; Marland, 1972). In the past decade, however, educators have offered several proposals for increasing the diversity of programs for the gifted and talented (see Boothe & Stanley, 2004, for one compendium of views). Some educators would make greater use of parent, teacher, and self-rating scales. Some would emphasize nonacademic talents—such as musical, athletic, and leadership abilities. Others call for the use of different measures—especially nonverbal ability tests. Although each of these proposals has something to offer, none directly addresses the central problem in identifying academic talent. All hope to find ways of measuring academic promise that will simultaneously reduce average differences between ethnic groups and identify those minority students who currently display or who are most likely to develop academic excellence. Unfortunately, just because a measure shows a smaller average difference between groups does not mean that it is a better tool for identifying academically talented students. Indeed, research has consistently shown that academic achievement in minority children is best predicted by the same cognitive and affective characteristics that predict academic achievement in majority children (Keith, 1999; Lohman, 2005). Therefore, although attempts to select students using other criteria may identify more minority students, many of the students identified—minority or majority—will not be the ones who currently display or who will someday develop academic excellence. The problem, then, is to find a way to identify and assist these children without compromising the ability of programs to serve children who already display high levels of academic and cognitive development. The goals of this monograph are (a) to explain why an aptitude approach to talent identification accomplishes these goals and (b) to illustrate how schools can implement this approach.

Identifying Academically Talented Minority Students

David F. Lohman
The University of Iowa
Iowa City, Iowa

EXECUTIVE SUMMARY

Poor and minority children in the U.S. are underrepresented in programs for academically talented students (Donovan & Cross, 2002). Attempts to increase the representation of these students, however, has been made difficult by recurring misconceptions about the nature of academic giftedness, the interpretation of measures of ability and achievement commonly used to identify gifted students, and the kinds of the educational programs that have developed to serve gifted students. I first discuss some of the misconceptions about giftedness that have thwarted efforts to identify academically talented minority students. Chief among these is the fallacy that intelligence tests can be constructed that measure innate ability. I then present an alternative model for identifying academically talented students that is grounded in modern theories of aptitude. In a nutshell, my argument is that (a) academic talent is best understood as aptitude for the kinds of expertise that can be developed through schooling; that (b) the primary aptitudes for academic learning are current knowledge and skill in a domain, the ability to reason in the symbol systems used to communicate new knowledge in the domain, interest in the domain, and persistence in the pursuit of excellence; and that (c) inferences about academic talent are most defensible when made by comparing a student's behavior to the behavior of other students who have had similar opportunities to acquire the knowledge and skills measured by the aptitude tests; however, (d) educational programming and placement should be based primarily on evidence of current accomplishments compared to all other students.

Aptitude refers to the degree of readiness to learn and to perform well in a particular situation or domain (Corno et al., 2002). Of the many personal resources that individuals bring to a situation, the few that assist them in performing well in that situation function as aptitudes. Previously acquired knowledge and skills in the domain of study are important aptitudes for all academic learning. For example, phonemic awareness is critical for learning to decode words; knowledge of algebra is critical for learning calculus. Reasoning abilities are the second major category of scholastic aptitudes. Instruction that requires students to discover relationships and make extrapolations places heavy demands on reasoning abilities, especially students' abilities to reason with verbal and quantitative concepts. Motivation, interest, and persistence are also important aptitudes for the attainment of academic excellence. Finally, aptitude cannot be assessed independently of the instructional and training programs in which students will be placed. For example, persistence under one set of learning conditions often differs from persistence under other learning conditions. This means that there must be "congruence between the criteria used in the identification process and the goals

and types of services that constitute the day-to-day activities that students will pursue" (Renzulli, 2005, p. 11).

The word *aptitude* is often used interchangeably with words such as *ability*, *talent*, and *potential*. However, *aptitude* is a more general term than *ability*: It includes those competencies called achievements as well. *Aptitude* is also a more inclusive term than *talent*. Academic talent commonly refers only or primarily to the cognitive aspects of aptitude, thereby excluding the broader range of motivational, temperamental, and other characteristics required for the development of expertise. *Aptitude* is easier to define and measure than *potential*. Potential is often taken to mean something like the level of competence that individuals might achieve if reared in environments that were perfectly attuned to their needs (Cronbach, 1972). When interpreted in this way, there is no way to measure the construct.

An aptitude approach to identification of academically talented minority students has several implications for practice. The first implication is that to identify the right students one must measure the right aptitudes. In general, the abilities, achievements, interests, and motivational characteristics that predict success in different academic domains are the same for students from all ethnic backgrounds (Keith, 1999; Lohman, 2005). For example, reasoning abilities in the symbol systems used to communicate new knowledge are critical aptitudes for all academic learning. Verbal reasoning ability is the single best predictor of academic achievement across a wide range of fields of study. Verbal reasoning ability in the language of instruction is often a better predictor of academic success for English Language Learners (ELL) students than for monolingual students. Other verbal skills—such as verbal fluency and production abilities—have also emerged as important aptitudes for academic success from first grade (Scott & Delgado, 2005) to college (Sternberg et al., 2004). Therefore, if one hopes to identify the most academically talented minority students, one should measure these and other characteristics that students must be able to call upon when learning. Practically, this means going well beyond the sort of nonverbal reasoning tests that can be administered to all children simply because it is assumed that all can be compared to a common norm group. Not to measure the appropriate aptitudes means that one will not identify many minority students who either currently exhibit the highest competence in particular academic domains or who are the students most likely to develop it. For example, selecting students on the basis of their performance on a nonverbal reasoning test actually excludes many of the most academically capable Black students.

The second implication is that one must compare students' scores on tests, rating scales, and other measures to the proper norm groups when making different kinds inferences from those scores. When the academic competence of students is estimated, the primary reference groups are given by the performance of all other children in the district, the state, and the nation. Blacks and Whites, males and females, rich and poor are all held to the same standards.

However, inferences about aptitude require more nuanced judgments. We say that a person has aptitude or talent for something if he or she learns in a few trials what

others learn in many trials. The critical variable, then, is number of opportunities the person has had to acquire the knowledge or to learn the skill. When abilities are developed over many thousands of trials, opportunity to learn must be approximated in other ways. For abilities developed through formal schooling, the amount of schooling provides a convenient yardstick. Grade norms on achievement tests use this standard. Inferences about aspects of the student's mathematics aptitude made from such scores presume that the student's educational experiences may be indexed by her year and month in school. Changing the norm group by even a few months can substantially change estimates of a student's aptitude, especially for young children. For abilities developed through more general interactions with the culture, *age* is used as the yardstick. Inferences about academic aptitude made from performance on ability tests presume that the student's experiences in the culture are similar to those of others who have lived in that culture for the same number of years and months. If the student has not lived or participated in the culture for all of that time, then inferences about her aptitudes made by comparing her scores to those of others of the same chronological age will be inappropriate. Concretely, when one is making inferences about aptitude to reason verbally in the English language, the score of a third grade ELL child should be compared with the scores of other third grade ELL children in the same school or district who have had roughly similar opportunities to learn English. This is not difficult to do, as I show in the final section of this monograph. Attempts to use a common, national norm group to estimate the academic talents of all students lead either to the use of tests and/or other procedures that are inferior measures of academic aptitude or to the identification of very few minority students. It is better to get a noisy estimate of aptitude using the right norm group than a more precise estimate using the wrong norm group.

The third implication is that students of the same age who are inferred to have talent in a particular area often have markedly different instructional needs. All students need instruction that is geared to their current levels of accomplishment. When students have had different opportunities to learn, however, instruction that is appropriate for one will often be inappropriate for the other. An undifferentiated label such as "gifted" does not usefully guide educational programming for a group that contains a mix of students with uneven discrepancies between accomplishment and aptitude for learning in different domains. One child may need from instruction several years in advance of her classmates; another may need more rapid coverage of the material being learned by her classmates; a third may need instruction at some level between these extremes or other kinds of support and encouragement. In addition to intensive instruction in the domain, minority students often need assistance in acquiring a vision of themselves as developing scholars. This can be particularly difficult when there is little social support for—and even disparagement of—academic excellence, especially from peers.

Historically, programs for the talented and gifted (TAG) were designed to serve students who were much more homogeneous in levels of achievement. Only those students who exceeded a common standard on tests of academic ability and/or achievement were targeted for special assistance. However, such policies are increasingly being challenged by minority students, their parents, and educational professionals. Programs that endeavor to serve both academically advanced students and

those academically talented students who display less stellar achievement thus face difficult choices. On the one hand, continuing with present policies preserves the ability of programs to serve the small and relatively homogeneous population of advanced students. This especially is the case for programs with classes specifically dedicated to serving the gifted. Simply adding minority students to these classes who are not prepared for the level of instruction that they will encounter serves no one well. However, continuing present practices may result in the increasing marginalization of such programs within the educational system. If this occurs, the already meager funding for TAG programs in many school districts is likely to be even further curtailed. On the other hand, rethinking the goals of TAG programs and the range of students and services that they provide could move programs in the opposite direction. Programs that target academically talented students for special assistance could be viewed as central to the school's mission if they broaden the range of services they offer to assist not only those who already exhibit high achievement, but also those who need more assistance in converting their superior academic talents into academic excellence.

References

- Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., & Talbert, J. E. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J. (1972). Judging how well a test measures: New concepts, new analyses. In L. J. Cronbach & P. Drenth (Eds.), *Mental tests and cultural adaptation* (pp. 413-427). The Hague, Netherlands: Mouton.
- Donovan, M. S., & Cross, C. T. (Eds.). (2002). *Minority students in special and gifted education*. Washington, DC: National Academy Press.
- 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.
- Lohman, D. F. (2005). The role of nonverbal ability tests in the identification of academically gifted students: An aptitude perspective. *Gifted Child Quarterly*, *49*, 111-138.
- Marland, S. P. (1972). *Education of the gifted and talented. Vol. 1. Report to the Congress of the United States by the U.S. Commissioner of Education*. Washington, DC: U.S. Government Printing Office.
- 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, University of Connecticut.
- Scott, M. S., & Delgado, C. F. (2005). Identifying cognitively gifted minority students in preschool. *Gifted Child Quarterly*, *49*, 199-210.
- Sternberg, R. J., The Rainbow Project Collaborators, and the University of Michigan Business School Project Collaborators. (2004). Theory-based admissions testing for a new millennium. *Educational Psychologist*, *39*, 185-198.

Table of Contents

ABSTRACT	vii
EXECUTIVE SUMMARY	ix
Introduction	1
Nonverbal Tests	2
Why Nonverbal Tests Will Not Solve the Problem	2
Uses of Nonverbal Tests in Screening for Gifted Students	5
Defining Giftedness	7
How Many Forms of Giftedness?	7
Is Giftedness Innate?	8
Giftedness as a Category Label	10
Giftedness as Relative to the Norm Group	12
Abilities as Achievements	16
The Jangle Fallacy	16
A Theory of Personal Theories about Ability and Achievement	17
Level 1. Naïve Nominalism or "Things Are What They Seem to Be"	18
Level 2. Ability and Achievement Test Seen as Exchangeable	18
Level 3. Complications Everywhere!	19
Level 4. Systems Theories	22
An Aptitude Theory of Academic Talent	22
A Definition of Aptitude	23
Effects of Context	24
Inferring Aptitudes	24
Scholastic Aptitudes	25
Common Methodological Pitfalls	27
The Non-exchangeability of Measures	27
<i>"And", "Or", or "Average?"</i>	30
Long-term Predictions of Achievement	31
Combining Scores From Different Tests	33
Identifying Academically Talented Minority Students	34
Prediction of Achievement for Minority Students	34
Assumptions About Growth	35
Judging Test Bias by Mean Differences Rather Than by Predictive Validity	36
The Need for Within-group Comparisons	38
A Sample Data Set	39
Caveat: Selection as an Ill-structured Problem	46
Suggestions for Policy	47

Table of Contents (continued)

References

51

List of Tables

Table 1.	Means (and <i>SDs</i>) for CogAT6 SAS Scores, by Ethnicity for a Random Sample of 300 Students	39
Table 2.	Students With the 50 Highest Scores on CogAT Verbal Battery + ITBS Reading Total	42
Table 3.	Students With the 50 Highest Scores on CogAT Quantitative Battery + ITBS Math Total	45

List of Figures

Figure 1.	Heritability of WISC Full Scale IQ Scores by SES	9
Figure 2.	Distribution Obtained When Selecting Only Examinees With High Scores	11
Figure 3.	Example of the Flynn Effect	12
Figure 4.	Proportion of Cases Exceeding the Same Cut Score on Two Tests, by the Correlation Between the Tests	28
Figure 5.	Schematic Plots of the "And," "Or," and "Average" Selection Rules	31
Figure 6.	Students in the Top 7% of the ITBS Mathematics Total Distribution at Grade 4 Who Were Still in the Top 7% at Grades 6 and 9	32
Figure 7.	Relative Contributions of CogAT Verbal, Quantitative, and Nonverbal Scores to the Prediction of ITBS Reading Achievement for White and Hispanic Students	35
Figure 8.	Plots of ITBS Reading Versus CogAT Verbal Reasoning, by Ethnicity	41

Identifying Academically Talented Minority Students

David F. Lohman
The University of Iowa
Iowa City, Iowa

Introduction

The purpose of this monograph is to present, explain, and then illustrate an aptitude approach for identifying academically talented minority students. Although grounded in over 80 years of research on the conceptualization and measurement of aptitude, the approach draws most heavily on the work of the late Richard Snow (see Corno et al., 2002). Sections of the monograph are adapted from four other papers that I have recently published on this topic. Most of these papers use data from Form 6 of the Cognitive Abilities Test (CogAT), which I now coauthor with Elizabeth Hagen. The first paper (Lohman, 2005) presents evidence that shows why nonverbal tests (such as the Nonverbal Battery of CogAT) should be measures of last resort—not first resort—when identifying academic giftedness. The second paper (Lohman, in press-a) gives an earlier effort to describe procedures for identifying academically talented minority students. The concepts and procedures offered here represent a more recent summary of my evolving understanding of this topic. The third paper (Lohman, in press-b) explains why ability tests are best viewed, not as measures of innate ability, but as achievement tests of a special sort. The fourth paper (Lohman & Korb, in press) explores why the majority of children who excel on measures of ability or achievement in primary school do not excel on similar measures administered in high school. Those who have time are encouraged to look at the original documents for a less abbreviated discussion of the points made here. All are available on my website, <http://faculty.education.uiowa.edu/dlohman>.

I first summarize some common misconceptions about the nature and measurement of giftedness that have thwarted efforts to identify academically talented minority students. Then I outline a method for identifying talented students that not only increases the proportion of underrepresented students who are served, but is more likely to identify those minority students who either currently display excellence in a domain or who are likely to develop it. Although the approach can apply to giftedness in any area, the discussion focuses on academic talent. In a nutshell, my argument is that (a) academic talent is best understood as aptitude for the kinds of expertise that can be developed through schooling; that (b) the primary aptitudes for academic learning are current knowledge and skill in a domain, the ability to reason in the symbol systems used to communicate new knowledge in the domain, interest in the domain, and persistence in the pursuit of excellence; and that (c) inferences about academic talent are most defensible when made by comparing a student's behavior to the behavior of other students who have had similar opportunities to acquire the knowledge and skills measured by the aptitude tests; however, (d) educational programming and placement should be based primarily on evidence of current academic accomplishments.

Joan's Dilemma

The parents were not like those Joan typically had to placate. In fact, the mother seemed genuinely perplexed as to why her daughter did not qualify for gifted services. Nicole had always been precocious. Last year, as a fourth grader, she had advanced to the honors algebra class at the local junior high. She was an inveterate reader who had already written several short stories that had amazed her teachers. But this year Nicole did not qualify, and her mother wanted to know why.

Joan had been the director of the TAG program now for almost five years. Last year she had initiated a new identification system in an effort to increase the representation of minority students in the program. Joan had attended several meetings in which it was explained that traditional procedures for measuring ability and achievement were biased against those who did not speak English as a first language or who did not have access to the hidden curriculum assumed by the tests. A new test was purchased that promised to provide a culture-fair way to measure ability. The presenter said that the test predicted academic achievement as well as other ability tests and, more impressively, would identify equal proportions of White, Black, and Hispanic students. It seemed almost too good to be true, but the presenter was very persuasive.

Joan and many of her colleagues in other school districts were eager to do a better job. Later that year she had administered the new test to every student who was nominated for admission to the TAG program. About half of those who had previously been included in the program no longer made the cut. The new group included a few more Hispanic students, but actually had fewer Black students. Further, many of the students who were identified were not doing that well in class. Joan remembered the sinking feeling when she first saw the test scores. Had she done something wrong? Had the teachers nominated the right children? Perhaps all along she had mistaken academic smarts for real giftedness. This was all very troubling.

Joan's dilemma is one that I have heard in one form or another from many directors of programs for the talented and gifted (TAG). All are good people who wanted to increase the diversity of their program for the gifted. They expected that their new identification procedures would discover more talented minority students. However, they did not anticipate that these procedures would exclude many children that they had formerly identified as gifted. And they did not expect that so many of the students—minority and majority—would now be classified as "nonverbal" learners who needed special programs.

Nonverbal Tests

Why Nonverbal Tests Will Not Solve the Problem

Since the earliest days of testing, nonverbal ability tests have been used to make inferences about the reasoning abilities of those who were not fluent speakers of the dominant language of the culture. Nonverbal tasks have long formed an important part of both individual intelligence tests (especially the Wechsler scales) and group ability tests such as the Otis-Lennon School Ability Test (Otis & Lennon, 1997) and the Lorge-

Thorndike Intelligence Tests (Lorge, Thorndike, & Hagen, 1964). We have now accumulated an enormous amount of information about nonverbal tests. I have been actively involved in this work for over 30 years. In a recent article (Lohman, 2005), I summarized some of the evidence for group-administered figural reasoning tests such as the Progressive Matrices (Raven, Court, & Raven, 1983) and the Nonverbal Battery of the Cognitive Abilities Test (Lohman & Hagen, 2001a). Briefly, here is some of what we know:

1. To call a test *nonverbal* is to make a statement about the observable characteristics of the items that are presented and the responses that are required. It is not—or at least should not be—a claim about the cognitive processes typically employed to solve items. Specifically, many nonverbal tests either require or elicit verbal processes in examinees. For example, the ability to label stimuli and the rules that describe the relations among them is critical for success on figural reasoning tests such as the Progressive Matrices (Raven et al., 1983). Therefore, language is important, even on nonverbal tests.
2. Some of the same students will be identified as gifted by nonverbal reasoning tests as would be identified by verbal or quantitative reasoning tests. Approximately 30 to 40% of U.S. students from all ethnic backgrounds show approximately the same level of reasoning abilities on nonverbal tests as on verbal or quantitative reasoning tests. This means that nonverbal tests will identify these students as well as either verbal or quantitative reasoning tests. However, the majority of students of all ethnic backgrounds do not score at the same level across the three major reasoning abilities. In all ethnic groups there are people who excel in verbal reasoning, others who excel in quantitative reasoning, and others who excel in figural-spatial reasoning. In all ethnic groups, those who show particular strengths in verbal and/or quantitative reasoning ability are the *most likely* to succeed academically. However, most of these students will *not* be identified by nonverbal tests. Indeed, students from all ethnic backgrounds who show a relative *strength* in nonverbal reasoning are actually *less likely* to succeed in school than those with similar levels of verbal and quantitative reasoning ability and a relative *weakness* in the nonverbal area. (See point 3.)
3. Nonverbal reasoning tests—especially those that require some form of analogical reasoning—are good measures of the general (*g*) factor that is common to all cognitive tasks. However, about half of the variation in nonverbal test scores is task specific. This means that differences between students on such tests are as likely to be caused by something specific to the test as by differences in their general cognitive ability. Oddly, many who are skilled in psychometrics ignore this when discussing test validity.

4. A test predicts success in learning to the extent that it samples the abilities required for successful learning. Success in school depends primarily on students' abilities to (a) acquire concepts through listening and reading and to (b) communicate with others through speaking and writing. Secondly, it requires the same sorts of abilities for quantitative thinking. Tests that measure the abilities to reason verbally and quantitatively are therefore the best predictors of subsequent success in school. This holds for all students, regardless of ethnicity.
5. Spatial abilities have been shown to be important additional aptitudes for success in mathematics and engineering that can usefully supplement measures of verbal and quantitative reasoning in predicting vocational choice (Shea, Lubinski, & Benbow, 2001). However, nonverbal reasoning tests such as the Progressive Matrices Test (Raven et al., 1983), the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997), or the Nonverbal battery of CogAT (Lohman & Hagen, 2001a) measure figural reasoning abilities, not spatial abilities. Good tests of spatial ability typically require visualization and mental transformation of three-dimensional objects (Lohman, 1994). Such tests also show large sex differences that favor males. Neither the Progressive Matrices test nor the CogAT show significant sex differences (in fact, females score slightly higher than males on the CogAT Nonverbal Battery). Therefore, if the goal is to identify students who excel in visual-spatial abilities, one should administer tests of spatial abilities, not tests of figural reasoning ability. Moreover, the spatial and figural types of thinking that are emphasized on nonverbal tests are required infrequently in school. Although it would be a good thing to design educational programs that assist these students, it would seem prudent first to identify those who can succeed in the programs schools currently offer. The burden for pioneering such programs should not be placed on the backs of minority students.
6. An important appeal of nonverbal tests is that differences between the average scores of native and nonnative speakers of English are smaller on such tests than on verbally loaded tests. (The differences between nonverbal and quantitative reasoning tests depend on the verbal demands of the quantitative test.) Thus, if common norms are used, proportionately more English Language Learners (ELL) and bilingual students will be identified than if selection is based on a test of verbal reasoning. Unfortunately, these will generally *not* be the ELL students who show the highest achievement test scores, especially in domains that rely heavily on verbal skills. Thus, for the convenience of using common norms, one gets only a fraction of the most academically capable ELL students.
7. The situation for Black students is quite different. Differences between Black and White students are, if anything, larger on nonverbal tests than on verbal and quantitative reasoning tests. Indeed, the profile of higher

verbal and quantitative abilities with lower nonverbal reasoning abilities is more common among Black students than among Whites, Hispanics, or Asian-Americans. Approximately 20% of Black students show this score profile (Lohman, 2005). This means that screening students with a nonverbal test will actually *eliminate* many of the most academically capable Black students.

8. Nonverbal ability tests *appear* to measure something more innate than tests that use words or other obviously learned symbol systems. This makes them highly appealing to those who want to believe that tests can measure innate ability. However, the inference of innateness is not justified. Scores on nonverbal tests are as much the product of education and experience as are scores on other types of reasoning tests. Indeed, norms for such tests have changed dramatically in the past 50 years (see *Flynn Effect*, p. 12ff). There are no culture-free or culture fair tests (Anastasi & Urbina, 1997; Scarr, 1994).
9. Naglieri and Ford's (2003) claim that the NNAT identifies equal proportions of high-scoring White, Black, and Hispanic students was supported only after the data had been re-weighted to make this happen. Because the data were contrived, other investigators have not been able to replicate these findings, either with Black students (Shaunessy, Karnes, & Cobb, 2004; Stephens, Kiger, Karnes, & Whorton, 1999) or with Hispanic students (Lewis, 2001). Indeed, all of these investigations found that the NNAT identified fewer high-scoring minority students than other nonverbal ability tests.

Uses of Nonverbal Tests in Screening for Gifted Students

What, then, is the proper role for nonverbal ability tests in identifying students for acceleration or enrichment? Such tests do have a role to play in this process. But it is as a measure of last resort, not of first resort. Height and weight are positively correlated. We can predict weight from height, but only with much error. It is not fairer to measure height for everyone just because we find it difficult to measure the weight for some. Rather, we should use predicted weight only when we cannot actually weigh people. High scores on figural reasoning tests tell us that students can reason well about problems that make only the most elementary demands on their verbal and quantitative development. This is extremely useful when one must determine whether a child suffers from a general cognitive impairment. From the early form boards of Itard to the contemporary nonverbal ability tests like the Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998), nonverbal tests have assisted psychologists in making inferences about the abilities of children who have hearing or speech problems or who do not speak the language of the examiner.

But it is one thing to make inferences about basic mental competence and another to infer readiness to profit from advanced instruction. Absent information on verbal

skills in the language of instruction, scores on a nonverbal test tell little about whether children will succeed in classes conducted in Spanish, Japanese, or any other language. More importantly, even *within* the population of native speakers, those students with the highest nonverbal reasoning scores are usually *not* the students who are most likely to show high levels of achievement in the classroom. *Rather, those students with the highest academic achievement in specific domains and those who reason best in the symbol systems used to communicate new knowledge in those domains are the ones most likely to achieve at a higher level.*

Therefore, high nonverbal scores should qualify students for acceleration or enrichment *only if* the scores are accompanied by (a) evidence of reasonably high accomplishment in the academic domain in which accelerated instruction or enrichment is offered or by (b) evidence that the student's verbal or quantitative reasoning abilities are high *relative to other children who have had similar opportunities to develop these abilities*. Most schools have this evidence for achievement, and those that administer ability tests that appraise verbal and quantitative reasoning in addition to nonverbal reasoning have the corresponding evidence for ability as well. For many ELL students, mathematics achievement and/or quantitative reasoning abilities are often strong, even when compared to the achievements of non-ELL students. For Black students, on the other hand, low scores on the nonverbal reasoning test are relatively common among those students with strong verbal and quantitative reasoning abilities. Thus, less than stellar performance on a nonverbal test is even less informative for Black students than for other students.

Absent ancillary information on verbal or quantitative abilities and achievement, the odds are not good that one will identify many of the most academically capable students by using a nonverbal, figural reasoning test. High scores on a nonverbal test are thus a useful supplement. They sometimes add to the prediction of achievement—especially in the quantitative domains. This means that the student with high scores on both the nonverbal and quantitative tests is more likely to excel in mathematics than is the student with high scores on either measure alone. (For verbal domains, the nonverbal test adds little. In fact, it sometimes has a negative influence on the prediction of success in domains that require verbal fluency.) And because the average scores for ELL students are generally higher on nonverbal tests than their scores on tests with verbal content, the test scores can encourage students whose academic performance is not strong. *The critical point, however, is not to confuse a higher average nonverbal score with better assessment of the relevant aptitudes.* Put differently, the figural reasoning test may appear to reduce bias, but when used alone, it actually increases bias by failing to select those most likely to profit from instruction.

Defining Giftedness

Any serious attempt to address the issue of minority representation in programs for the gifted brings to the fore important questions about the nature of giftedness and the way it is identified. Indeed, how best to identify gifted children is one of the most persistent and controversial topics in the field of gifted education (VanTassel-Baska, 2000). Much of the controversy stems from different beliefs about the meaning of the term *gifted*. Should giftedness be restricted to academic domains, or should it include artistic, athletic, leadership, and other types of competence valued by society? In the academic domain, should it be based on evidence of superior academic accomplishments or on a measure of academic potential, such as IQ?

How Many Forms of Giftedness?

Although many espouse broadening the definition beyond traditional notions of IQ, considerable disagreement remains on which domains should be included. For example, should programs be devised to develop all of Gardner's (1983, 2003) intelligences? And if not, what is the principle that ranks one higher than another? For those who adopt an ability-centered approach, should selection be based only on general ability (*g*), or should it include the eight broad-group factors at the second level of the Cattell-Horn-Carroll (CHC) theory (McGrew, 2005)? If the definition includes group factors, are all eight factors equally important? If not, then why not? What are the principles that lead one to develop one set of abilities and not another? And if selection will be based on *g*, how does one account for the fact that different measures of *g* select different students who have quite different likelihoods of displaying or developing particular kinds of academic expertise?

Even a cursory consideration of these questions reveals that we are not interested in ability for ability's sake, but rather in ability *for* something. We are not interested in identifying bright kids in order to congratulate them on their choice of parents or some other happenstance of nature or nurture. Rather, the goal is to identify those children who either currently display or who are likely to develop excellence in the sorts of things we teach in our schools. Identifying such students is a more manageable problem than trying to measure the hundreds of ways in which people differ and then creating programs that are uniquely tailored to each kind of exceptionality. Put differently, those who take an ability-centered approach to the identification of giftedness have no basis other than parsimony for designating one ability as more important than another ability. For example, it is only when we add the criterion of utility that general crystallized abilities become much more important than general spatial or general memory abilities in the identification of academic giftedness. Crystallized abilities better predict school achievement, even though general crystallized, spatial, and memory abilities have equal stature in the modern theories of human abilities. Additionally, unlike the aptitude paradigm that I advocate, the ability-centered approach to giftedness offers no principled way for incorporating motivation, creativity, or any of the other factors we may think important into the selection process. Indeed, the MENSA Society International is the example *par excellence* of the ability-centered approach to the identification of giftedness.

Is Giftedness Innate?

The common conception of giftedness emphasizes the importance of innate ability. The innate ability model helps account for the fact that exceptional children are typically exceptional from a very young age. Although many of these children come from families that have encouraged their development, this encouragement is often as much a response to the child's unusual abilities as a cause of them.

Genetic factors are clearly important in accounting for individual differences in achievement and ability. The statistic that estimates the proportion of genetic variation in a trait for a particular group of individuals in a particular range of environments is called a *heritability coefficient*. For example, in the U.S., heritability for height is about .85. This means that about 85% of the variation in children's heights can be predicted by knowing their parent's heights. In countries where there is a much greater variation in nutrition, the heritability drops to about .60. The proportion of genetic variance is smaller primarily because the environmental variation is greater. Traits that are not affected by genetic factors would have a heritability of zero.

Two facts are commonly overlooked in discussions of the extent to which genetic factors explain or account for individual differences in intelligence. First, the contribution of heredity is typically as high for measures of achievement as it is for measures of ability. The distinction between ability and achievement tests cannot be made on the basis of the contribution of heredity to individual differences in the scores. (See *The Jangle Fallacy*, p. 16.) Second, estimates of heritability vary substantially across cohorts of people who share a common genetic heritage. For example, Sundet, Tambs, Magnus, and Berg (1988) computed heritability for the ability test administered to all 19-year-old Norwegian men at the time that they become eligible for military service. The yearly data go back to 1931. Heritability coefficients followed a gentle sine curve that repeatedly peaked and dipped across decades, with a high value of about .80 and a low of about .20. Heritability is thus not a number but a range of numbers.

Sometimes we can identify the environmental factors that produce these changes in heritability coefficients. For example, it has long been known that the contribution of heredity to IQ scores varies markedly across social classes (Jensen, 1981). However, the issue is difficult to study because within ethnic groups in the U.S., social class is in significant measure a function of ability. Estimating the unique contributions of heredity and environment requires not only much data on low socioeconomic status (SES) families but also the use of complex statistical techniques to disentangle the variables.

A recent study by Turkheimer, Haley, Waldron, D'Onofrio, and Gottesman (2003) excelled on both counts. Figure 1 is adapted from their report. The solid line in the leftmost panel of Figure 1 shows how heritability for WISC IQ scores at age seven varied as a function of socioeconomic status for 319 twins. The best estimate of heritability for the lowest SES children (those with SES scores between 0 and 20) was near zero. This means that individual differences in the WISC IQ scores of these twins were almost entirely due to variations in their environments. For the highest SES children (those with

SES scores between 80 and 100), the estimated heritability was between 80 and 90% of the variance. Therefore, the contribution of the environment—both that portion shared by both children and that portion unique to each child—was vastly more important for low SES children. The environmental variance shared by the twins is shown in panel C and the environmental variance unique to each twin is shown in panel E.

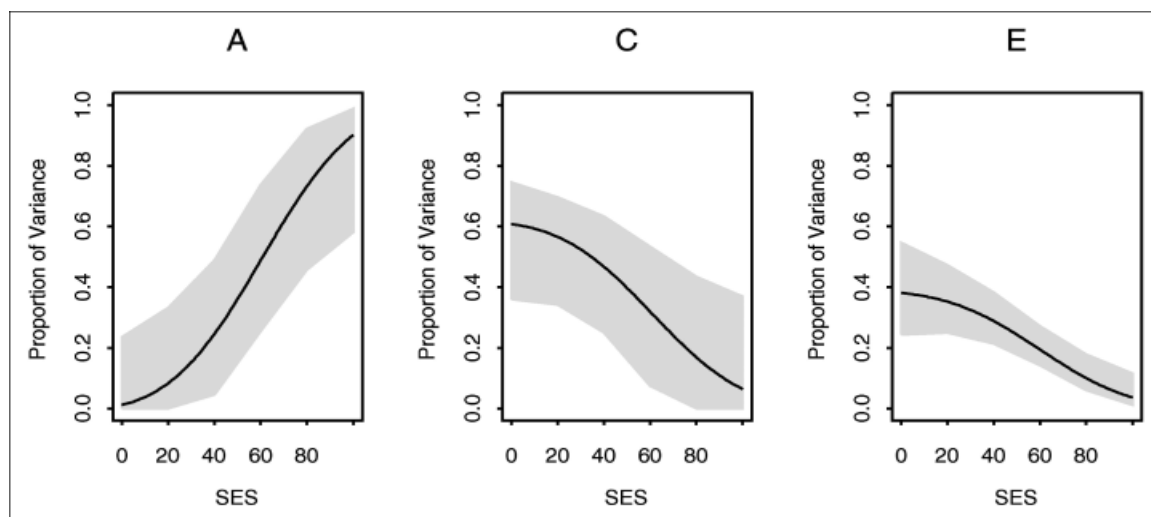


Figure 1. Proportion of variance in *WISC* Full Scale IQ at age 7 accounted for by additive genetic (Panel A), shared environmental (Panel C), and non-shared environmental components (Panel E) as a function of socioeconomic status (SES) for $N = 319$ twin pairs; 43% White, 54% Black, mostly low SES families (Turkheimer et al., 2003). (Reprinted with permission.)

This does not mean that ability tests are biased against low SES children. Most do a good job of estimating current levels of cognitive development. What Figure 1 shows is that the contribution of genetic factors to differences in performance varies as a function of the child's opportunity to develop the abilities estimated by the test.

Policies for identifying gifted students make explicit the policy maker's assumptions about the relative importance of biology and experience on the development and expression of abilities. Selection policies that emphasize the importance of biological factors include (a) basing identification on tests that are assumed to measure innate ability (especially IQ) rather than on achievement or on some combination of ability and achievement, (b) identifying gifted students early in primary school—even kindergarten—but not systematically thereafter, or (c) making inferences about academic potential by comparing the scores of all students to the same national norm group. Conversely, selection policies that emphasize the role of experience tend to (a) emphasize measures of accomplishment or academic achievement, usually within particular domains; (b) aim for yearly identification and reassessment of students, with services matched to the current development of the child; and (c) estimate students' aptitude by

comparing their performance not only to all other students but to other students who have had similar opportunities to acquire the knowledge and skills measured by the test.

Selection policies also make concrete assumptions about what tests measure. Indeed, part of the controversy about identification practices stems from misunderstandings about the limitations of tests and other scales. Many identification procedures erroneously treat test scores as if (a) they were error free, (b) all tests measured the intended construct with equal fidelity, (c) all tests measured the same thing throughout the score scale, and (d) the norms of the different tests were equally good. Ignoring these limitations of tests can seriously compromise the identification process.

Giftedness as a Category Label

Words both direct and mislead our thinking. The act of naming something enables communication with others. But it can also reinforce the perverse human tendency to misrepresent a characteristic that varies continuously. Thus, we speak of *learning disabled* or *gifted* students as if these labels represented discrete categories rather than arbitrary portions of continuously varying score distributions. If two well-respected tests give different ability scores for the same child, one saying that the child is gifted and the other reporting a lower score, many would dismiss one outcome—usually the lower score. But such disagreement between tests is the rule, not the exception. Suppose we define *gifted* as scoring in the top 3% of the distribution of intelligence. We administer two of the best individual intelligence tests—the Stanford-Binet V and the WISC-IV—to each student. Full-scale IQ scores on these tests correlate about $r = .84$ (Roid, 2003). This means, however, that only about half of the students who score in the top 3% on one test will also score in the top 3% on the second test (Lohman & Korb, in press). If the interval between test administrations is longer than a few weeks, then even fewer would merit the label *gifted* on both tests.

This is not what most test users expect, in part, because when we think about gifted children, we tend to envision those children who most clearly exemplify the category. These will generally be those with the most extreme scores. And although the scores for these children are also likely to differ across tests, both scores are likely to fall above the cut score. For other students, we have an unfailing tendency to focus on those scores that are consistent across occasions because they confirm our expectations of consistency.

Confirmation bias is widespread in human reasoning (Nickerson, 1998). However, no matter where we set the cut score on the upper tail of the score distribution, many more students will be near that cut score than far above it. This is shown in Figure 2. When the scores of these students regress towards the mean on retest—and on average they always regress—many will fall below the cut, and the scores of an equally large group of students who previously failed to make the cut will now rise above it. Regression to the mean is inevitable whenever the two sets of scores are not perfectly correlated. Indeed, it is another way of saying that two variables are not perfectly correlated. The lower the correlation or the higher the initial score, the greater the

regression.¹ Keeping track of everyone—not just the cases that stand out clearly in our minds—helps us combat the confirmation and typological biases in our thinking.

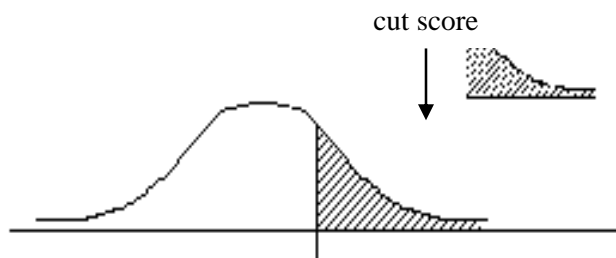


Figure 2. How selecting examinees with high scores produces an odd-shaped score distribution with many individuals just above the cut score. Many of the examinees in the left half of this new distribution will score below the cut if given an alternative test.

Even those who understand that the boundaries between "gifted" and "not gifted" are arbitrary often assume that category membership would remain constant if we had perfectly reliable measures. This is not true. Longitudinal studies of ability and achievement show that the majority of students who would be classified as gifted one year would not be so classified a few years later, even if we somehow obtained error-free scores on the ability test (Humphreys & Davey, 1988). This is because (a) cognitive abilities develop at different rates in different children and (b) the sources of individual differences change as one moves up the developmental scale.

The critical mistake here is to assume that ability is fixed, not constantly developing. It ignores the fact that to maintain a particular rank, a child must not only get better each year but must improve at the same rate as others who had the same initial score. Using status scores such as percentile ranks (or derivatives such as IQs) masks this year-to-year growth. If the same dimension were labeled "language development" rather than "giftedness," then we would expect to find some whose development was unusual at one point in time but not unusual at another. Speaking in sentences is unusual for a one year old. It is not unusual for a two year old. Those who identify gifted students can thus

¹ Here is a way to envision why regression occurs. Imagine the floor of a large auditorium in which students are sitting in rows of chairs arranged in the form of a normal (bell-shaped) distribution. The rows near the middle of the auditorium have many chairs, but the outside rows at the extreme right and left have only one chair. At a given signal, students rise and must find a new chair. They are allowed to move, on average, three rows in either direction. (How much movement is allowed is given by the correlation between the students' row numbers before and after the move.) Those near the middle will have no trouble finding a chair in the rows to their right or left. But for those near the extreme right, there will be many more chairs to their left than to their right. For every person who moves to the right, two or three will find places by moving toward the middle. Thus, when everyone is reseated, those who were seated near the extremes will, on average, have moved closer to the middle.

inadvertently become gatekeepers for the precocious rather than advocates of developmentally appropriate instruction for all students.

Giftedness as Relative to the Norm Group

Flynn Effect. Judgments of exceptionality depend on the norm group. Scores that are unusual in one cohort often are not unusual in another. For example, national norms for both ability and achievement tests have been changing for as long as we have been norming tests. Scores on ability tests have been rising at the rate of about three IQ points per decade since the 1920s. This increase is sometimes called the *Flynn Effect* after the researcher who first systematically documented it in many countries (see, e.g., Flynn, 1987). Figure 3 shows one estimate of this increase for the Binet and Wechsler tests.

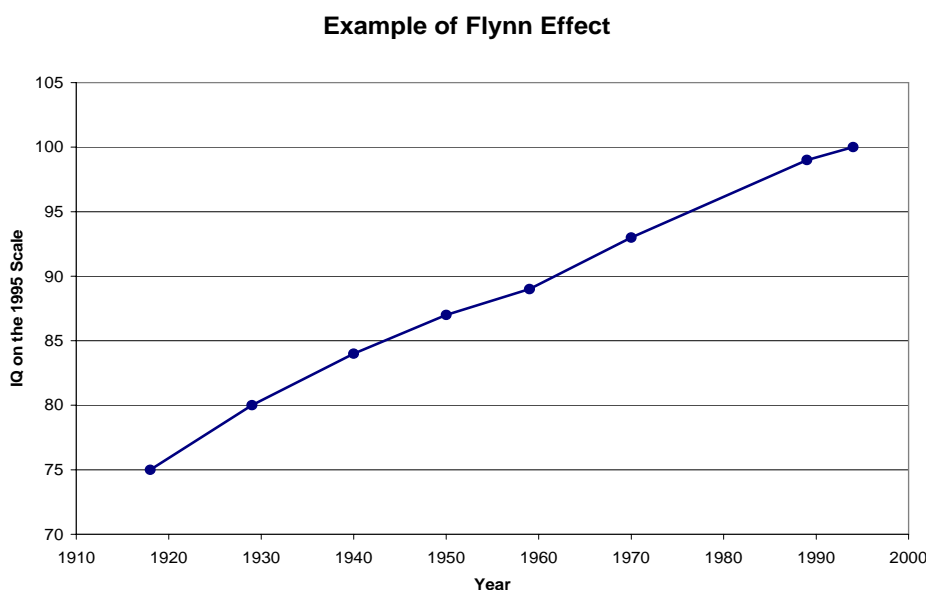


Figure 3. Gains in average IQ for the U.S. White population on the Binet and Wechsler tests from 1918 to 1996. Examinees who received an IQ of 100 in 1918 would have received an IQ of 75 for a similar performance in 1996 (Horgan, 1995).

Growth has been even larger on nonverbal tests such as the Progressive Matrices and was unabated in the most recent studies (see Raven, 2000). This is one reason that major ability and achievement tests are re-normed every 5 to 10 years. Even when two tests are normed in the same year, however, samples of examinees differ, and so norms for the two tests are not necessarily the same. Therefore, a selection rule that defines admission in terms of IQ or national percentile rank will not admit the same number of students in different years or when different tests are used, especially if the norms are not of equal recency and quality.

There are many examples of studies that demonstrate how differences in norms can confound efforts to identify gifted students. For example, Shaunessy et al. (2004) administered the Culture Fair Intelligence Test (CFIT; Cattell & Cattell, 1965), the Standard Progressive Matrices (Raven; Raven et al., 1983), and the NNAT (Naglieri, 1997) to 196 predominantly Black students in a low SES rural school district. Their goal was to see which test identified the most gifted students. To do this, they compared the number of students who fell in 5-point percentile bands on each test beginning at the 80th percentile.

The NNAT, which has the most recent norms, identified only three students as falling above the 80th age Percentile Rank (PR); the Raven, which has some normative data collected in the 1980s, identified 18 students; and the CFIT, which has the oldest and least defensible norms, identified 36 students.² There were two findings. First, the study showed that the NNAT did not identify equal proportions of White and Black students (cf. Naglieri & Ford, 2003). Second, it demonstrated the importance of the norming sample and how norms for tests have changed over the past 50 years.

Even if the tests were all normed on the same population, the common practice of accepting the highest score across several different tests that measure the same ability is fundamentally misguided. It assumes that the highest score in a set of different test scores best estimates a student's ability. This is not true. The best estimate is given by the average of these scores (see pp. 38ff).

The Importance of Local Norms. The Flynn Effect demonstrates that judgments about exceptionality depend on the national norms that are used to interpret scores. More students will obtain high scores when older norms are used than when more recent norms are used. However, differences between schools in the same state are many times greater than differences between cohorts of students in different decades. This is important because the need for special services depends not so much on a student's standing relative to age- or grade-mates nationally, but on the student's standing relative to the other students in the class. Talent searches and district-wide programs that recruit students from different schools need the common standard of national, state, or district norms. National norms also provide critical information on a student's relative standing on the different abilities measured by the test. Individual schools, however, rarely replicate the nation in their distribution of ability or achievement. In about 5% of the schools in the nation, the *average* student scores at the 95th percentile on the Iowa Test of Basic Skills. Surely the students in these classes are quite capable. But it is unlikely that a student who scores at the 98th national percentile in such a class will be as mismatched with the common curriculum as the student who scores at the 98th national percentile in a class in which the typical student scores at the 50th percentile on the ITBS. In short, although both national and local norms have important uses, decisions about identification and

² Although the CFIT norms are still in use, they are not recommended. They were based on convenience samples of U.S. and British students collected in the 1960s and were indefensible when new (see Tannenbaum, 1965). National norms for the Raven have never been collected, so users are advised to collect their own norms.

acceleration are often best made using local norms. Many publishers offer local norms when the school or district tests all children in a particular grade.

Age Norms. Judgments about academic potential often assume a different cohort than either a national or local grade cohort. Suppose that we discover that the student whose achievement is exceptional is actually a year older than the other students in the class. For example, some parents hold their child out of school for a year to give him an advantage in physical and cognitive development over his classmates. Although instruction should be geared to the child's achievement, would one still consider the child "gifted?" Conversely, suppose a child is considerably younger than her classmates or has attended school irregularly. Should her scores be compared with others in the same grade when estimating her ability to learn?

One of the primary differences between ability and achievement tests is that ability tests report scores relative to age-mates.³ Ability is an inference about rate of learning given equal opportunity to learn. We use age as a yardstick in measuring ability because it is a useful surrogate for "total amount of experience in the culture." If the abilities are those that can be developed in the course of everyday interactions with the culture, then comparisons to one's age cohort provide important information. If the abilities can be developed through school experiences, then comparisons with those who have had similar amounts of education (i.e., grade norms) are also helpful. However, even small differences in the choice of age cohort (e.g., 6 years 0 months versus 6 years 10 months) can make a large difference in whether a particular score is considered exceptional. If the child was ill for several months or lived in the culture for only half of her life, would norms based on her age cohort be most appropriate for inferences about her ability to learn?

Subgroup Norms. For those whose experiences differ markedly from the norm, aptitudes need to be judged relative to a different cohort. *Always, the preferred comparison group would be those who have had roughly similar opportunities to acquire the abilities sampled by the test.* Concretely, one should look at the performance of the ELL child relative to other ELL children who have had roughly similar amounts of exposure to English. The "fair" procedure of comparing all to the same age or grade group regardless of their experience is equivalent to the "fair" procedure of comparing all children to a common score distribution, regardless of age. Furthermore, as shown below (see pp. 51ff), one need not develop rigorous norms tables or compare the child only to the handful of others who have had similar experiences. Rough classifications (such as ELL versus native speakers) go a long way to correcting the problem.

Assessments in Other Languages. Should a test also be administered in the child's other language(s)? If the goal is to assess the full extent of a child's cognitive competence, then the test(s) needs to be aligned not only to the language(s) but also to the culture(s) of the child. This is more commonly a problem when making judgments about cognitive impairments than about readiness to profit from a more rapidly paced

³ For individually administered ability tests, this is the only norm group. Group-administered tests such as CogAT report both age and grade norms.

instruction, for example. In such situations, language can introduce construct-irrelevant variance into the testing situation. One can reduce this impact by presenting problems that require verbal reasoning, but that use only pictures. This allows the child to use any language when attempting items. For example, one can present verbal analogies using pictures rather than words. Unfortunately, many concepts cannot be readily displayed in pictures and even those that can are often difficult to interpret (e.g., is it a tomato or an orange?). Further, pictures are not culturally neutral. In spite of these limitations, such tests provide a more comprehensive picture of the child's intellectual development than those that only require reasoning with geometric shapes.

However, instruction is always conducted in one or more languages. Knowing the child's competence in another language is at best a helpful predictor of the level of competence the child might someday attain in the language of instruction. Concretely, if a child has excellent oral language abilities in Spanish, we would predict the attainment of at least above-average skills in English. Spanish listening, speaking, reading, and writing abilities will function as direct aptitudes for classroom learning, however, only if the instructional program allows or requires the child to use them. In those situations in which English is either one of the languages of instruction (or the only language of instruction), students' performance relative to others who have had roughly similar opportunities and experiences in acquiring English should be estimated. To exclude such abilities from the assessment will significantly underrepresent the set of aptitudes needed for learning. Assessments in the language of instruction provide an important frame of reference for making judgments about the likelihood that if given proper assistance, the child will someday attain academic excellence in an English-speaking educational system. In this way, an aptitude perspective thus helps clarify those situations in which assessments in a second language would be helpful or even necessary.

Scaling Effects. Raw scores (i.e., number correct) on most standardized tests are first converted to scale scores. IQ scores are simply age percentile ranks (PRs) of the distributions these scale scores. An IQ of 100 always translates to an age PR of 50. The PR equivalent of other IQ scores depends on the standard deviation that is imposed on the scores. Different procedures for constructing score scales will produce different raw score to scale score conversions, and thus will result in different IQ scores. For example, changes in the scaling of the Stanford-Binet between Form L-M and the fourth and fifth editions dramatically reduced the number of extremely high IQ scores that were reported (Ruf, 2003).

Conclusions. Judgments about exceptionality depend importantly on the norm group that is used. Whether a particular score is considered exceptional also depends on how the norms were derived, how the test scores were mapped onto a score scale, and how the scores will be interpreted. The child who is considered gifted when compared to others in his class may not be considered gifted when compared to others in the nation, to others who are the same age, to those who were tested a few months earlier, to examinees of the same age who were tested a decade or two later, or to those who have had more experience in the culture of the assessment. Those who do not understand the relativity of norms—especially on ability tests—miss the easiest and most effective way to identify

those minority students who are most likely to develop academic excellence. It is important to measure the right abilities; but it is equally important to compare students' scores to the right norm groups.

Abilities as Achievements

Consider two fifth grade students. One shows academic accomplishments well in advance of her peers but an ability test score that is good but not exceptional. The other shows the opposite pattern: a high ability test score but only moderate academic accomplishments. Which student is most clearly academically gifted? Many in the field of gifted education would pick the child with the higher ability test score because they believe that the ability test is a better measure of innate ability, whereas measures of academic accomplishments (such as achievement test scores, science projects, etc.) measure "school learning," not real giftedness.

There are two problems with this belief. First, it assumes educational accomplishment depends solely on those abilities sampled by the ability test. Other abilities or personal attributes and contextual factors such as the content of the class or the method of instruction are assumed to exercise minor influence on learning. But this is not the case. Although general reasoning and problem solving abilities are single most important resource for learning, they are only one of many personal and contextual factors that must be brought to bear in order to develop competence in a domain. An aptitude perspective requires that one understand these other contextual and personal factors. This should be done at the outset, not as an afterthought. The aptitude perspective also sidesteps fruitless debates about the nature of intelligence. Reasoning abilities are but one of several aptitudes required for the development of academic excellence. They are surely an important measure, and absent any information about prior accomplishment, provide the best prediction of academic success. But they are a predictor, not the criterion. The criterion is academic excellence in any of its many forms. Confusing the predictor with the criterion would be like confusing physical fitness with skill in playing basketball, tennis, or any other sport. Many who have the requisite physical fitness required by sport do not have other physical, motivational, or personality characteristics needed to develop high levels of competence in it.

The second reason many people give precedence to IQ over evidence of accomplishment is that they, like many before them, have an intuitive theory of ability as innate. This theory comes not from some conspirators bent on misleading the public. Rather, like intuitive theories in science, it comes from their everyday experiences. Understanding the alternative view requires understanding something about the development of personal theories of ability and achievement.

The Jangle Fallacy

Ability tests are best understood as a particular type of achievement test. However, since the earliest days of mental testing, both psychologists and educators have

struggled to attain this understanding. In a book published in 1927, Truman Kelley called attention to something he dubbed the "jangle fallacy." Kelley was the lead author of the first edition of the Stanford Achievement Test. As a statistician of some repute, Kelley was bothered by the way people treated scores on his achievement test and various intelligence tests as if they measured independent constructs. He knew that the overlap in individual differences on the two types of tests was enormous. For the tests he was using, Kelley estimated that about 90% of the "true" or systematic variation in general intelligence was shared by composite measures of academic achievement.

The culprit, he said, was language. Because different words—"intelligence" and "achievement"—were used to describe the constructs measured by the two types of tests, people treated intelligence and achievement test scores as if they were in fact distinct. Kelley coined the term "jangle fallacy" to describe this tendency to treat terms that sound different as if they really signified different concepts. "Intelligence" and "achievement" sound as though they are (or should be) different things. Therefore, we treat tests so labeled as if they measured different things.

Kelley was surely at least half right. Language indeed abets our propensity to view concepts that differ only in degree as if they differ in kind (Nickerson, 2004). But language not only shapes thought, it is in turn shaped by the concepts and belief systems we have constructed and wish to express. The larger problem, then, may be the beliefs themselves.

A Theory of Personal Theories About Ability and Achievement

Like most fallacies in educational measurement, the jangle fallacy seems to recur in every generation. In large measure, this is because most of us who study abilities and achievements go through a similar progression in our beliefs about the nature and nurture of intellectual competence. As in most such developmental schemes, virtually everyone begins with a version of the simplest theory. Why some move on while others remain committed to a particular belief is not always clear. One factor that seems to matter is a willingness to consider—and even to seek out—evidence that contradicts one's current views. Openness to new perspectives is difficult if one has a vested interest in preserving the current belief. Cognitive styles may also matter. A fondness for the sort of sharp categories that typify simpler theories may make it difficult to move to fuzzier worlds in which there is more grey than black and white.

There are several clear examples of this progression in beliefs about intelligence and achievement (see Lohman, in press-b). For many, the transition to a new understanding of ability comes after realizing that what intelligence tests measure is very much the product of education and experience. J. M. Hunt recounts such a transition. His book, *Intelligence and Experience* (1961), summarizes research on the effects of experience on the development of intelligence. In trying to explain why challenges to his belief in a fixed intelligence were so unnerving, Hunt appealed to Festinger's (1957) theory of cognitive dissonance:

In his own professional life history, the writer finds in himself some evidence of [cognitive dissonance]. So long as he was professionally identified with the testing function, it was highly comforting to believe that the characteristics tested were fixed in individuals. Evidence hinting that these characteristics were not fixed produced intense dissonance, for it threatened his belief in fixity and the adequacy of his professional function as well. (pp. 14-15)

There are many other examples of this transition in beliefs about the relationship between conceptions of intelligence and achievement. For most theorists, the changes are less well documented than in Hunt's case. But the products of these changes are clear, even though the transitions themselves are usually less transparent. The development of beliefs about the constructs of ability and achievement commonly fall in one of the following four levels.

Level 1. Naïve Nominalism or "Things Are What They Seem to Be"

The person at this level believes that ability tests measure (or ought to measure) innate potential. This means that scores on an ability test should not be influenced by culture, education, personal experience, or motivation. Similarly, achievement tests measure (or ought to measure) only knowledge and skills learned in school. Performing better on an ability test than on an achievement test is interpreted as "not living up to one's potential." Virtually everyone starts with this understanding. They retain this belief until confronted with evidence that challenges it. Then they either progress to the next level or engage in various repair strategies to preserve their Level 1 beliefs.

Level 2. Ability and Achievement Tests Seen as Exchangeable

The person at this level has encountered evidence of the overlap between measures of intelligence and achievement. This evidence may come from statistics that show high correlations between ability and achievement tests. Less formally, it may come from an inspection of ability and achievement tests that shows similarities in the content and structure of items on the two types of tests.

Reactions to this knowledge take several forms. Some look at the overlap and conclude that all tests measure general ability (Spearman, 1923; Jensen, 1998). Others look at the same data and say that the overlap is mostly the product of learning (Ferguson, 1956; Humphreys, 1981; Thorndike, Bregman, Cobb, & Woodyard, 1926). Although both explanations account for the overlap between ability and achievement tests, neither accounts for the differences between them. A popular solution that attends to both the similarity and difference is to envision a continuum in which tasks vary by their novelty. The more achievement-like tasks are at the low-novelty end of the continuum, whereas the more ability-like tasks are at the high-novelty end. A continuum like this may be found in the writings of Stern (1914), Thorndike et al. (1926), Anastasi (1937), Cattell (1943; 1963), Cronbach (1970), Snow (1980), and Sternberg (1985), to name a few. Placing both types of tasks on the same continuum recognizes their commonality. Placing them at opposite ends of the continuum also recognizes their uniqueness.

Interpretations of this continuum vary. For example, in his original proposal of the theory of fluid (*Gf*) and crystallized (*Gc*) abilities, Cattell (1943) emphasized the equality of the two intelligences that he had distinguished. However, in later versions of the theory (Cattell, 1963, 1971), fluid ability was interpreted as something like the true, innate intelligence of the individual that, when invested in experience, produced a particular constellation of crystallized abilities. Although Horn (see, e.g., Horn & Noll, 1997), Snow (1980), and others have repeatedly disputed this interpretation, it is the one that is most often presented in elementary textbooks that discuss the theory of fluid and crystallized abilities. Among other things, the view of *Gf* as the measure of innate intelligence fails to explain why genetic factors are as important for *Gc* as for *Gf* (Cronbach, 1976).

The attempt to interpret fluid reasoning abilities as the real intelligence is perhaps better understood as an attempt to preserve Level 1 beliefs about innate ability in the face of Level 2 evidence that contradicts these beliefs. There are other common examples of this tendency. Most people who recognize that current tests of ability and achievement overlap also assume that we could construct tests that would greatly reduce or eliminate the overlap. An example is the hope expressed by many of us in the 1970s that we could construct better ability tests by directly measuring the higher-level cognitive processes that people used when solving items on ability tests or performing other complex tasks (e.g., Hunt, Frost, & Lunneborg, 1973; Snow, 1978; Sternberg, 1977). A more extreme version of this view is being championed today by those who claim that only "nonverbal" tests can measure abilities. In this view, tests that measure reasoning abilities should not be contaminated by content or skills that would influence performance on an achievement test. Some advocate the use of nonverbal tests such as the Progressive Matrices (Raven et al., 1983) or the Nonverbal Battery of the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001a). Others consider even these tests contaminated because the directions are given orally and thus use words. Advocates of this view argue that achievement tests ought not to measure anything that could be labeled "ability" (Naglieri & Ford, 2005).

Level 3. Complications Everywhere!

Those who get beyond the idea that ability and achievement are separate—or presently may be made separate—confront a long list of further complications. Many who enter this swamp seem never to emerge. Rather, like marooned naturalists, they contentedly explore the wonders of a series of island domains, each of which is sufficiently complex to occupy a competent research team for their entire careers. Examples include:

The effects of culture on cognition. Beginning in the 1920s, some theorists noted that the very concept of intelligence is rooted in culture (see Anastasi, 1937; Degler, 1991). A culture-free measure of intelligence is thus something of an oxymoron. Similarly, what counts as achievement varies across cultures and eras. Factual knowledge, spelling, and computation skills are less valued today than one hundred years ago. Independent thinking and problem solving are generally more highly valued today.

The effects of education, practice, and training on abilities. All abilities—from those required by the simplest reaction-time task to the most complex problem-solving task—respond to practice and training. Near the end of her career, Anastasi (1980) observed that much confusion could be avoided if the term "ability" would always be prefaced by the adjective "developed." Similarly, Snow observed that intelligence is not only education's most important raw material but also its most important product (Snow & Yalow, 1982). Of course, this does not rule out a substantial role for biological factors in individual differences in abilities at any point in the sequence of their development.

The effects of knowledge on thinking. Just as people too glibly speak of the distinction between "ability" and "achievement," many also speak too glibly about the separation of cognitive processes and the knowledge on which those processes act. One of the most important discoveries about human cognition is the extent to which thinking is bound to the objects of thought (Greeno, Collins, & Resnick, 1996). Put differently, there are no "information-free" cognitive processes. Reasoning does not exist as a module in the brain that can be applied like a tool to different problems (Lawson, 2004). Rather, how well we reason depends on how much we know. Language has particularly powerful effects on the development of thought, from the acquisition of simple perceptual concepts to complex assemblies of knowledge and skill that require many years to acquire.

The unity of the ability/achievement space. If all abilities are achievements, and all thinking is rooted in knowledge, then it makes little sense to talk about abilities and achievements as if they were qualitatively different things (Snow, 1980). Rather, many who study individual differences see a single space of developed competencies or abilities (Carroll, 1993; Cronbach, 1990; Horn & Noll, 1997; Humphreys, 1981). Some of these abilities are developed primarily through formal schooling, others through out-of-school experiences common to most children in a culture, and yet others through experiences that are unique to the individual. Different tests sample from different regions of that space of competencies. Achievement tests sample a broad range of knowledge and skills acquired primarily in school-like activities. Ability tests such as Otis-Lennon or CogAT emphasize reasoning abilities that are required by and developed through experiences both in and out of school. Most individually administered ability tests, such as the Stanford-Binet V (Roid, 2003) and especially the Woodcock-Johnson III (Woodcock, McGrew, & Mather, 2001), sample a much broader array of abilities such as Comprehension-Knowledge, Long-term retrieval, Visual-Spatial ability, and Short-Term memory. Put another way, although reasoning depends on knowledge, what is known includes much more than those associations that facilitate reasoning. Reasoning abilities are thus a subset of the much larger domain of developed competencies.

The multidimensionality of the unified ability space. For a very long time we have known that ability is a multidimensional, not unidimensional concept. Most theorists agree that the 70+ abilities that have been identified can be organized in a hierarchy: a *g* factor at the highest level, seven or more broad group abilities at the next level, and 50-87 primary abilities at the base (Carroll, 1993; Horn & Blankson, 2005; McGrew, 2005). This theory is much less comforting to those who wish to make simple

comparisons between ability and achievement than the Cattell (1963) theory of fluid and crystallized abilities. Fluid reasoning ability now includes verbal, quantitative, and figural aspects (Carroll, 1993); crystallized abilities no longer include quantitative achievements (McGrew, 2005). There is no easy division of abilities into the two camps of ability and achievement constructs.

Is Gc the real intelligence? There has long been a bias among researchers that fluid intelligence (*Gf*) represents the real, biologically determined intelligence, whereas crystallized intelligence (*Gc*) better represents the products of investing this biological intelligence in particular experiences. Although there is some evidence that this may indeed be the case with very young children, thereafter "*Gc* may precede and do more to determine *Gf* than the reverse" (Horn & Blankson, 2005, p. 64; see also Lohman, 1993). *Gf* may also not be as central to models of abilities as many have believed. In his last published paper, Carroll (2003) reported a reanalysis of the Woodcock-Johnson (Revised) norm data. One of the major purposes of the analysis was to test Gustafsson's (1988) hypothesis that $Gf = g$. The analysis, however, showed that the *Gf* factor was much less important than expected. The best measures of *g* were vocabulary and mathematical problem-solving tests.

The impact of affect and volition on cognition. Aristotle distinguished between cognition, affection, and conation—or knowing, feeling, and willing. Modern research on cognition shows that thinking is deeply enmeshed with affect. Interest (or disinterest), surprise (or boredom), enjoyment (or disgust) moderate what we remember about a topic, how deeply we think about it, and how long we will persist in thinking about it. Similarly, the choices that we make as we embark on a task, or when we first encounter difficulty or distraction, also influence the success of our efforts. Put differently, one cannot estimate how well people think unless they are willing to try their best. Even then, they will generally do better if the topic interests them and if they feel that they are having success at it. Further, the knowledge and skills that they assemble both reflect and feed into interest. There is no way to separate the measurement of ability from motivation or feeling.

The effects of experience on brain structures. As with other dichotomies, the simple distinction between biology (or genotype) and experience does not survive close inspection. For example, even if one could somehow measure the neuronal connections of the neonate's brain, the measure would not describe her biological structure for long. We now know that the brain is changed by experience. Extensive experience in a domain effects substantial changes in the structure of the brain and the way it processes information (Nelson, 1999). At a molar level, this means that the biological contribution to individual differences in ability is moderated by the quality of the environment in which the child is raised. Heritability is substantial for children in high SES families but much lower for children in low SES families (see Figure 1). Further, as Cronbach (1976) noted, virtually any statement made about the heritability of tests of general ability would apply with equal force to measures of academic achievement. This does not mean that anyone can do anything. A naïve environmentalism is surely more misleading than a naïve nativism. What it means is that, just as the neonate must grow physically to keep

up with his peers, so too must he grow cognitively. The experiences that feed this growth determine subsequent cognitive status even more surely than nutritional intake determines subsequent physical growth.

Level 4. Systems Theories

Given the scope and complexity of research on cognition, it is not surprising that very few scholars are able to envision theories or paradigms that integrate these diverse themes into a coherent whole. Two of the most impressive efforts are those of Robert Sternberg and Richard Snow.

Most educators are familiar with Sternberg's (1985) triarchic theory, so it will not be summarized here. Snow's theory is less well known but in many ways is easier to apply to the problems educators face (for an introduction, see Corno et al., 2002). The theory concerns how we might best design instruction to meet the needs of different learners (Cronbach & Snow, 1977). Turned around, it addresses the fundamental question of readiness to learn from a particular set of educational activities (Snow & Lohman, 1984; Snow, 1992, 1994). The central construct is that of aptitude, by which Snow meant readiness to learn and perform well in a particular situation or domain. Aptitudes for learning are therefore tied both to what must be learned (i.e., what kind of expertise do we aim to develop?) and to the learning context (i.e., how are students expected to learn?). Students who will have a difficult time acquiring one type of expertise (e.g., mastering algebra) may have less difficulty acquiring expertise in another domain (e.g., creative writing). Those who might have difficulty succeeding under one instructional arrangement (e.g., large lecture class) might succeed more readily under another (e.g., computer-assisted instruction). Deciding which students are most likely to develop a particular type of expertise thus begins with a careful analysis of what constitutes expertise in the domain and how it is developed. Next, one looks at the demands and opportunities of the different educational paths offered for those who wish to develop expertise in the domain. What must students know and be able to do to succeed in each alternative route to the attainment of expertise? The theory thus turns the question of "intelligence" on its head. One begins not with a catalog of the person's scores on a given set of ability dimensions but rather with a clear statement of where one wants to go (What kind of expertise?) and of the paths (What kinds of instruction?) that will be available.

An Aptitude Theory of Academic Talent

An aptitude approach to understanding academic talent is thus very much concerned with abilities but in a different way than in most theories of giftedness. The focus is on *all* of the aptitudes that must be brought to bear to accomplish something. In particular, the goal is to identify those children who either currently display academic excellence and are most likely to continue to display it, and those children who show less exceptional levels of accomplishment but are likely to develop it. Identifying such students is a much more tractable problem than identifying all the ways in which children

differ and then creating programs that uniquely fit each need. The first point, then, is that academic giftedness is best understood in terms of aptitude to acquire the knowledge and skills taught in schools that lead to forms of expertise that are valued by society. We are interested in ability tests only because they help identify those who may someday become excellent engineers, scientists, writers, etc. In other words, we are interested in abilities because they are indicants of aptitude. They are not the only indicants but one important class of indicants.

A Definition of Aptitude

So, what exactly do we mean by *aptitude*? Although often rooted in biological predispositions, it is not something that is fixed at birth. School achievements commonly function as aptitudes—for example, reading skills are important aptitudes for school learning. Indeed, aptitude encompasses much more than cognitive constructs such as ability or achievement. Persistence is an important aptitude in the attainment of expertise. Also, aptitudes are not necessarily positive. Some people have a propensity to experience or to cause accidents; others to lie; and still others to be unsociable or even hostile. The intuitive appeal of theories of emotional intelligence is rooted in the common observation that a productive and happy life requires more than abstract intelligence. Finally, and most importantly, the term *aptitude* is not a descriptor of a person that is somehow independent of context or circumstance. Indeed, *defining the situation or context is part of defining the aptitude*. Changing the context changes in small or large measure the personal characteristics that influence success in that context. Aptitude is inextricably linked to context.

Consider formal schooling. Students approach new educational tasks with a repertoire of knowledge, skills, attitudes, values, motivations, and other propensities developed and tuned through life experiences to date. Formal schooling may be conceptualized as an organized series of situations that sometimes demand, sometimes evoke, or sometimes merely afford the use of these characteristics. Of the many characteristics that influence a person's behavior, only a small set aid goal attainment in a particular situation. These are called aptitudes. Formally, then, aptitude refers to *the degree of readiness to learn and to perform well in a particular situation or domain* (Corno et al., 2002). Thus, of the many characteristics that individuals bring to a situation, the few that assist them in performing well in that situation function as aptitudes. Those that impede their performance function as inaptitudes. Examples of characteristics that commonly function as academic aptitudes include the ability to comprehend instructions, to manage one's time, to use previously acquired knowledge appropriately, to make good inferences and generalizations, and to manage one's emotions. Examples of characteristics that function as inaptitudes include impulsivity, high levels of test anxiety, or prior learning that interferes with the acquisition of new concepts and skills.

Effects of Context

The same situation that assists one student can thwart goal attainment in another. For example, discovery-oriented or constructivist approaches generally succeed better with more able learners while more didactic approaches may work better with less able learners (Cronbach & Snow, 1977; Snow & Yalow, 1982). Less-structured learning situations afford the use of the able students' superior reasoning abilities, which function as aptitudes. However, anxious students often perform poorly in relatively unstructured situations (Peterson, 1977). Thus, the same situation that affords the use of reasoning abilities can also evoke anxiety. Recent efforts to understand how individuals behave in academic contexts have emphasized the importance of traits that together produce the outcomes that we observe (Ackerman, 2003). Lubinski and Benbow (2000) have also argued for greater attention to diversity in the needs of academically gifted students. Indeed, gifted students will vary as much from each other on those dimensions least correlated with *g* as students in the general population.

Understanding which characteristics of individuals are likely to function as aptitudes begins with a careful examination of the demands and affordances of target tasks and the contexts in which they must be performed. This is what we mean when we say that defining the situation is part of defining the aptitude (Snow & Lohman, 1984). The affordances of an environment are what it offers, makes likely, or makes useful. Placing chairs in a circle affords discussion; placing them in rows affords attending to someone at the front of the room. Discovery learning often affords the use of reasoning abilities; direct instruction often does not. Unless we define the context clearly, we are left with distal measures that capture only some of the aptitudes needed for success. This is why *g*-like measures of ability correlate imperfectly with success in any particular school task, especially when students are allowed a choice over what they study and how they might go about it. On the other hand, averaging across learning situations and outcome measures obscures the impact of the particular abilities and magnifies the relative importance of *g*.

Inferring Aptitudes

Aptitude is commonly inferred in two ways. In the first, aptitude is estimated from the speed with which the individual learns the task itself. Aptitude for a task is inferred retrospectively when a student learns something from a few exposures that other students learn only after much practice. When available, this is the most unambiguous evidence of aptitude for learning something. Indeed, the concept of aptitude was initially introduced to help explain the enormous variation in learning rates exhibited by individuals who seemed similar in other respects (Bingham, 1937).

In the second way, we attempt to identify other tasks that require similar cognitive or affective processes and measure the individual's facility on those tasks (Carroll, 1974). Because these measures only predict success on the task, they will more often error in identifying those students who will excel in learning the task itself. For example, phonemic awareness skills that facilitate early reading in Spanish for Hispanic students

also facilitate early reading in English for these students (Lindsey, Manis, & Bailey, 2003). To estimate the probability that Spanish-speaking students will learn to read English, one can measure their phonemic awareness skills in Spanish. Similarly, dance instructors screen potential students by evaluating their body proportions, ability to turn their feet outwards, and ability to emulate physical movements (Subotnik & Jarvin, 2005). Although none of these characteristics requires the performance of a dance routine, all are considered important aptitudes for acquiring dance skills.

Scholastic Aptitudes

The most important requirement of most academic tasks is domain knowledge and skill (Glaser, 1992). Measures of prior knowledge and skill are therefore usually the best predictors of success in academic environments, especially when new learning depends heavily on old learning. Measures of current knowledge and skill include on-grade-level and above-grade-level achievement tests and well-validated performance assessments such as rankings in debate contests, art exhibitions, and science fairs. Performance assessments that supplement achievement tests offer the most new information if they require the production of multiple essays, speeches, drawings, or science experiments, rather than the evaluation of essays, speeches, drawings, or science experiments produced by others (Rodriguez, 2003). The more closely measures of accomplishment sample critical aspects of emerging expertise in the domain, the better they will capture aptitude for learning in that domain.

Inventories of conceptual and factual knowledge in a domain can provide critical information on this aspect of academic development. These are frequently overlooked, often because there is no easy way to rank all children on the same dimension. Most achievement tests—especially those designed for elementary school children—contain relatively little content knowledge. However, studies of the development of expertise show that to develop competence in an area of inquiry, students must construct rich networks of well-organized factual and conceptual knowledge (Bransford, Brown, & Cocking, 2000). This knowledge tends to be quite localized, especially when learning is self-directed. Bright children assemble vast amounts of knowledge about specific topics that are at best represented only superficially on achievement tests. An achievement test that is designed to be fair to all children can hardly be expected to reveal much about the specialized knowledge a student has acquired. In this respect, the dilemma that confronts those who would assess gifted children is the same dilemma that has stymied those who investigate adult intelligence. Hunt (2000) suggests that we might do a better job if the metaphor that guided the construction of the assessment were to conduct an inventory rather than a survey.

Short-term educational decisions should therefore rely primarily on evidence of current accomplishment in a domain. Other aptitudes enter the picture, though, with each step one takes into the future. For example, given the same type of instruction, continued improvement in a domain requires interest or at least dogged persistence. More commonly, continued success requires a new mix of abilities: Algebra requires some skills not needed in arithmetic; critical reading requires skills not needed in beginning

reading. Teachers, teaching methods, and classroom dynamics also change over time, each requiring, eliciting, or affording the use of somewhat different personal characteristics. Indeed, in most disciplines, the development of expertise requires mastery of new and, in some cases, qualitatively different tasks at different stages. Sometimes the critical factor is not only what is required for success but also what is allowed or elicited by the new context that might create a stumbling block for the student. For example, in moving from a structured to a less structured environment, a student may flounder because he is anxious or is unable to schedule his time. Indeed, I sometimes think that the attainment of expertise has as much to do with inaptitudes as aptitudes.

It should be no surprise, then, that the second most important set of personal characteristics for academic learning are the ability to go beyond the information given; to make inferences and deductions; and to see patterns, rules, and instances of the familiar in the unfamiliar. The ability to reason well in the symbol system(s) used to communicate new knowledge is critical for success in learning. Academic learning relies heavily on reasoning (a) with words and about the concepts that they signify and (b) with quantitative symbols and the concepts that they signify. Thus, the critical reasoning abilities for all students (minority and majority) are verbal and quantitative. Figural reasoning abilities are less important and show lower correlations with school achievement (Lohman, 2005).

Therefore, if the goal is to identify those students who are most likely to show high levels of future achievement, both current achievement and domain-specific reasoning abilities need to be measured. Analyses of the CogAT-ITBS data (Lohman & Korb, in press) suggest that the two should be weighted approximately equally, although the relative importance of prior achievement and abstract reasoning abilities will depend on the demands and affordances of the instructional environment and on the age and experience of the learner. In general, prior achievement is more important when new learning is like the learning sampled on the achievement test. This is commonly the case when the interval between old and new learning is short. With longer time intervals between tests or when content changes abruptly (as from arithmetic to algebra), then reasoning abilities become more important (Lohman & Korb, in press). Novices typically rely more on knowledge-lean reasoning abilities than do domain experts. Because children are universal novices, their reasoning abilities are more important in the identification of academic talent, whereas evidence of domain-specific accomplishments is relatively more important for adolescents.

Learning requires more than prior knowledge and good reasoning abilities. This is because learning is never a purely rational activity. Whether a child persists in thinking about something depends on affective and motivational factors. Sometimes affective engagement can be elicited by parents, teachers, and coaches. But more commonly high levels of engagement are better understood as a resonance or attunement between the child and the activity or domain. This fascination can be short-lived or enduring. Either way, interest is a critical and easily measured aptitude for learning or performing well. Interest inventories can be helpful, especially for adolescents (see

Lubinski, Benbow, & Ryan, 1995; Schmidt, Lubinski, & Benbow, 1998). For younger children, less formal methods may be used.

Finally, many of the more important characteristics of students that function as aptitudes for learning are best obtained through teacher ratings. For example, motivation, creativity, and expressive communication skills can be estimated through teacher ratings on the Scales for Rating the Behavioral Characteristics of Superior Students (Renzulli et al. 2002). Combining such information with information on interests, abilities, and achievements is still more art than science. Should teacher ratings be given greater weight than students' interests? And how should these measures be combined with estimates of ability and achievement? One of the most pressing needs in the field of gifted education is for longitudinal studies that would allow one to give empirically supported answers to such questions.

Common Methodological Pitfalls

Implementing a system for identifying academically talented children is fraught not only with conceptual problems, but also with statistical and psychometric traps. In this section, I discuss a few of the more common mistakes. Even seasoned professionals fall prey to some of these errors. Mostly this is because texts that are commonly used to teach correlational methods are written for psychologists and other researchers who hope to build theories about unobservable constructs. Errors of measurement and aspects of the assessment procedure that obfuscate relations among variables are removed at the outset. Those who use tests to select children cannot do this. When using test scores, one gets all that they measure—the construct of interest, a large dose of whatever is specific to the particular collection of tasks that are administered, and errors of measurement that inflate or depress the score that is obtained on a particular occasion under particular testing conditions. Further, problems that have only small effects, on average, can substantially impact affect extreme scores. And giftedness is all about extreme scores.

The Non-exchangeability of Measures

Some programs admit students if they obtain a sufficiently high score on any one of several measures of ability or achievement. For example, a student may be admitted if he obtains a sufficiently high score on either the WISC-IV or the Stanford-Binet V. Some take a high score on virtually any ability test. As previously discussed (see pp. 14ff), the first problem is that different tests may be normed on quite different populations, so using national norms (e.g., IQ scores) favors the test with the oldest and "easiest" norms. Suppose we eliminate this problem by administering both of these tests to everyone in the local population and, instead of using IQ scores or national percentile ranks, admitted those students who scored in the top 5% of the local distribution on either test. We accept a high score on either test because we know that the two tests are highly correlated. We assume that if tests are highly correlated, we would identify more or less the same individuals on either measure. However, this assumption is simply false. There

is much confusion about this in the educational literature, abetted in large measure by a misunderstanding of how to interpret correlations.

Suppose we administered two tests, plotted the scores for all students, and calculated the proportion that scored above a particular cut score (such as the top 5%). Next, we repeated the experiment using tests that showed different degrees of correlation. If we then plotted the proportion of students who were above various cut scores on both tests, as the correlation varied we would get the curves shown in Figure 4. The correlation between two tests is shown on the abscissa (x axis) of the figure. The correlation varies from $r = .5$ to $r = .975$. Plots for seven different cut scores are shown. These range from the top 1% (bottom curve) to the top 20% (top curve). For example, the bottom curve shows the proportion of students scoring in the top 1% of the distribution on one test who could also be in the top 1% of the distribution for the second test. The proportion ranges from only 13% when $r = .50$ to 76% when $r = .975$.

Figure 4 shows clearly that there will be considerable disagreement in classification among different tests unless the correlation is very high and the cut score is very low. Neither of these conditions typically applies. For example, the correlations among total scores (i.e., Full Scale IQ scores) on individually administered ability tests range from $r = .68$ to $r = .85$. A common criterion is scoring in the top 3% of the distribution. Figure 4 shows that this means only about one-third to one-half of the students would be expected to score in the top 3% on both tests. Scores for shorter tests (e.g., Verbal IQs) show lower correlations and would agree even less well.

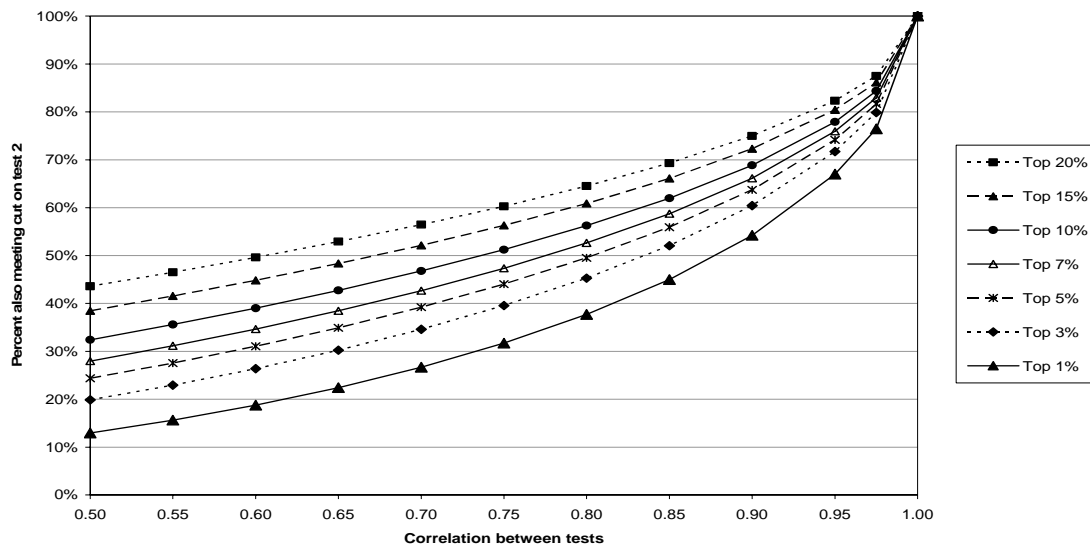


Figure 4. The graph shows the proportion of students exceeding a given cut score on one test that will also exceed the same cut score on a second test. Data show selected cut scores and correlations between the two tests.

Achievement constructs show similar effects. Consider reading abilities. What percentage of students who scored in the top 3% of the distribution of ITBS Reading Total scores (Reading Vocabulary plus Reading Comprehension) would we expect to identify using a series of other selection measures?⁴ These measures are ordered by their correlation with the ITBS Reading Total Score. They are:

1. **ITBS Composite.** Many schools use the Composite across all subtests of the *ITBS* to identify academically gifted children. But what percentage of the best readers would be missed using this score? Reading comprehension is not only a critical aptitude for success on other subtests of the *ITBS*, but the Reading Total Score also enters into the computation of the *ITBS* Composite (so there is a statistical confounding as well). The median within-grade correlation between the Reading Total and the Composite was $r = .91$ in this national sample.
2. **CogAT Verbal Battery.** Verbal reasoning abilities are critical in the acquisition of both reading comprehension skills and reading vocabulary. The average within-grade correlation between the CogAT Verbal Battery and the *ITBS* Reading Total was $r = .82$.
3. **CogAT Composite.** In addition to the three battery scores, CogAT reports a Composite. It is the best estimate of g on CogAT. The median correlation between the CogAT Composite and the *ITBS* Reading Total score was $r = .79$.
4. **CogAT Nonverbal Battery.** Some schools use nonverbal reasoning to identify gifted students. Although this is surely the most distal battery studied, its median correlation with the Reading Total was still substantial (median $r = .62$).

Surely we would expect to identify most of the top readers using the *ITBS* Composite. However, Figure 4 shows that we would probably only get about 60% of them. This is not what most people would expect for two variables that correlate $r = .91$. Using the CogAT Verbal Battery, we would identify 47% of the best readers. The CogAT Composite would get 44%. And the Nonverbal Battery would identify only 28% of the best readers. Clearly, different measures do not identify the same students in spite of the fact that the tests are highly correlated. Indeed, even very high correlations imply far less agreement between scores than most people think, especially for extreme scores.

Does this mean that any student who obtains a high score on a test should be considered gifted in the assessed domain? Not at all. A substantial part of the difference between the scores individuals obtain on two tests is due to the many influences we collectively call *errors of measurement*. If we were to administer a parallel form of the first test on a different occasion, then many students would not obtain high scores on both

⁴ The data come from the 2000 joint national standardization of Form A of the Iowa Tests of Basic Skills® (*ITBS*®; Hoover, Dunbar, & Frisbie, 2001) and Form 6 of the Cognitive Abilities Test™ (*CogAT*®; Lohman & Hagen, 2001a). These observed results are reported in greater detail in Lohman (in press-a). They show less agreement than predicted. This is because scores on achievement tests are often not normally distributed.

tests. The reliability coefficients of parallel forms rarely exceed $r = .90$. Even with this degree of reliability, Figure 4 shows that only 60% of the students who are tested would score in the top 3% on both tests.

The implications of these results, then, are that one should *never* base decisions about admission on a single score. Rather, one should use an **average** score of parallel forms of a well-chosen test administered at different times. How much difference will an average score make?

"And," "Or," or "Average"?

Sometimes a few syllables can make quite a difference. Those who remember the history of Western civilization may recall that the early Christian church split over a syllable in the Nicene Creed: was it to be *homoousios* (identity of essence) or *homoiousios* (similarity of essence)? Likewise, the shift between a rule that admits a student on the basis of a high score either on test 1 *or* on test 2 admits many more students than a rule that admits a student on the basis of a high score on both test 1 *and* test 2. The intermediate position—take the *average* of test 1 and test 2—admits yet a different group.

Figure 5 graphs the results of these three rules. Each panel shows a scatter plot of students' scores on the two tests. The left panel shows the students identified by the "and" rule, the center panel those identified by the "or" rule, and the right panel those identified by the "average" rule. Requiring students to score above a particular cut score on both tests 1 and 2 restricts the number of students who are identified. This is the effect of a two-stage screening process in which students must achieve a high score on the first test (e.g., a norm referenced achievement test) and then a high score on a second test (e.g., an individually administered ability test). Consider the case in which the cut score is set at the top 5% on both tests and the correlation between them is $r = .80$. Only about 50% of the students in the population who meet this criterion on one test will also meet it on the second test (see Figure 4 and the more extensive tables in Lohman & Korb, in press). This means that 50% of the 5% who met the criterion on test 1, or 2.5% of the total student population, will be admitted.

The "or" rule has quite different effects. Again, the percentage of students admitted is easily estimated. Test 1 admits 5% of the population. Test 2 also admits 5%, but half of these students were already admitted by the first test. Therefore, in all, 7.5% of the student population would be admitted. Changing the rule from "and" to "or" triples the number of students admitted from 2.5 to 7.5% of the population.

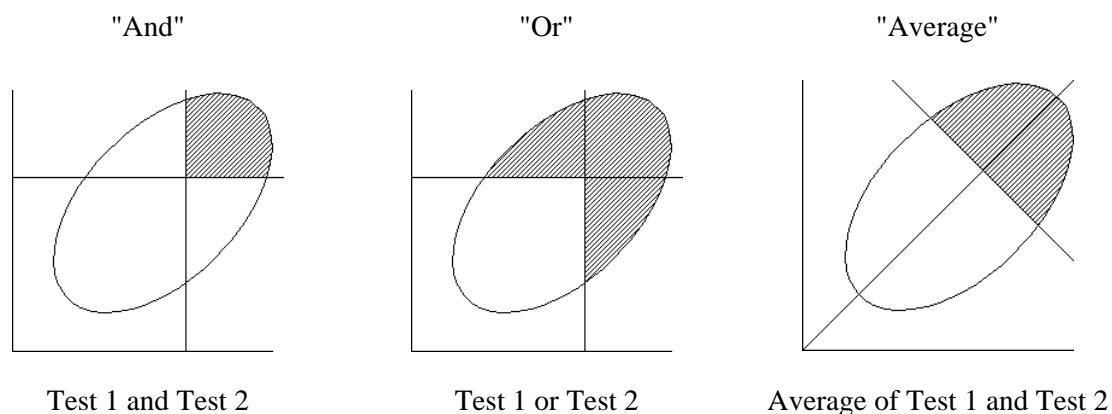


Figure 5. Plots of the effects of three rules: (a) high scores on test 1 and test 2; (b) high scores on test 1 or test 2; and (c) high scores on the average of test 1 and test 2.

The disjunctive "or" rule is most defensible if the two tests measure different constructs such as language arts and mathematics. One should seek to identify students who excel in either domain, not just those who excel in both domains. If both tests measure the same construct, however, the statistically optimal rule is neither "or" nor "and" but rather "average." As Figure 5 shows, the "average" rule will admit more students than the restrictive "and" rule and fewer than the liberal "or" rule. Essentially, students are admitted on the basis of where they fall on the 45° diagonal rather than on either the x axis or the y axis. Further, using two scores to estimate ability (as in either the "and" or the "average" rule) substantially reduces regression effects.

Long-term Predictions of Achievement

Although current achievement is a critical aspect of academic talent, it is also important to consider other characteristics that indicate readiness to continue to achieve at a high level. In addition to current accomplishments in a domain, readiness includes (a) reasoning ability in the major symbol systems used in that domain, (b) interest in that subject area, and (c) persistence in learning that domain. A recent study that I reported with Katrina Korb showed this (Lohman & Korb, in press). We did not have measures of interest or persistence, but we did have ITBS Survey Battery scores and CogAT reasoning scores in verbal, quantitative, and figural domains. Students ($N = 2,525$) in a large Midwestern school district were administered both ITBS and CogAT in grades 4, 6, and 9. Our analysis looked at the percentage of students whose achievement test scores were above the 93rd percentile in grade 4 who also had similarly high achievement scores in grades 6 and 9. Figure 6 shows the results for Mathematics. The solid line shows what happened when we identified students solely on the basis of their grade 4 math scores. Two years later, less than half scored above the 93rd percentile, and by grade 9 only about 40% remained in the group. These are typical regression effects.

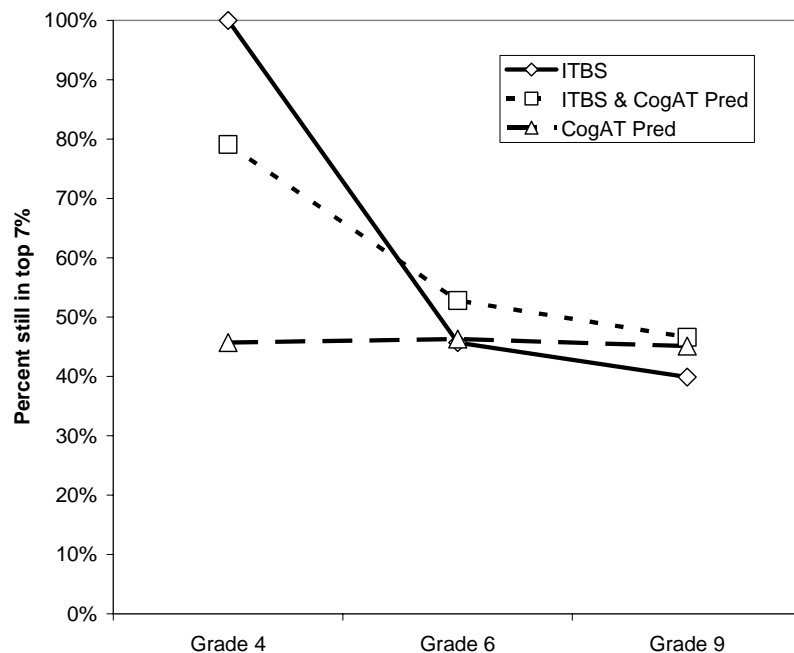


Figure 6. Students in the top 7% of the ITBS Mathematics Total distribution who were still in the top 7% of the distribution at grades 6 and 9 using three identification models at grade 4.

The heavy dashed line shows what happened when we used only grade 4 CogAT scores to select students. At grade 4, we missed many of the high scorers. But by grade 6, we identified the same proportion of high achievers as we did when using grade 4 achievement. By grade 9, the grade 4 CogAT scores identified more of those who were in the top 7% of the Math distribution than did grade 4 ITBS math scores.

Finally, the dotted line shows what happened when we combined grade 4 math and CogAT reasoning scores. This measure effected the most reasonable compromise. It identified most of the high achievers at grade 4, did better than either measure alone at grade 6, and performed slightly better than either measure alone at grade 9.

This study and others show that the best selection models combine both current achievement and reasoning abilities in the symbol systems used to communicate new knowledge in the domain. At grade 4 this makes good sense for both reading and mathematics. For reading, later accomplishments in literature, history, and similar domains depend on reading abilities just as they depend on reasoning abilities: as cognitive aptitudes used to acquire expertise. Similarly, mathematics at grade 4 is sufficiently different from later mathematical expertise (even algebra) to consider math achievement more as a measure of skills that are helpful for acquiring future expertise in mathematics. Multiple measures that estimate different aptitudes are also better than one measure that captures only some of the needed skills. Multiple measures—even of a single aptitude—also dramatically reduce the error of measurement in the selection model.

Combining Scores From Different Tests

Combining scores from different tests is thus almost always a better policy than using a single score. But how should scores be combined? With minimal justification, here are a few guidelines.

1. *If scores come from the same test (e.g., ITBS Reading Total scores for grades 3 and 4), average the scaled scores for the two administrations of the test.* For example, on the ITBS these are called Standard Scores. For CogAT they are called Universal Scale Scores. On CogAT and other ability tests, one can also average the Standard Age Scores (SAS).

2. *Expect that averaged scores will regress to the mean.* One cannot use norms tables to look up the percentile ranks of the averaged scaled scores in the same way that one looks up the percentile ranks of individual test scores. Similarly, averaged SAS scores will not have the same PR associated with each score as individual SAS scores. Therefore, base decisions on rank within the local group of all students' average scale scores, not on whether these average scores exceed a fixed percentile rank that applies only to individual scores.

3. *If scores come from tests that use different score scales, first put them on a score scale that has the same mean and standard deviation.* The easiest way to do this is to convert scaled scores to z -scores by computing the mean and standard deviation for each set of scaled scores. These z -scores have a mean of 0 and a standard deviation of 1. The standard deviation is one measure of the spread or dispersion of scores. By making the standard deviations the same, we insure that one variable does not overwhelm the other when they are combined. For example, to combine ITBS Reading Total Standard Score (SS) and CogAT Verbal Battery Universal Scale Score (USS) scores, first get the mean and standard deviation for each set of scores. Then compute $z(\text{reading})$ and $z(\text{Verbal Battery})$ using a function such as "standardize" in Microsoft Excel. Finally, add the two z -scores together to get a composite score that weights each of the component scores equally (see pp. 54ff of this monograph for an example).

4. *Generally avoid combining percentile ranks (PRs).* Use scaled scores instead. PR scores make only crude distinctions at the top of the score scale. A difference of 1 PR point may mean a difference of many points in scaled scores. This is another way of saying that the relationship between PR scores and scale scores is not linear.

5. *Base the weights assigned to different tests on research.* Although equal weights work well when combining ability and achievement test scores, one would not want to give equal weight to scores that are either less reliable or that have weaker relationships with outcome measures. For example, even though interest and persistence are critical, measures of these constructs are much less reliable and show at best moderate correlations with outcomes. It would not be appropriate to weight them the same as measures of achievement or ability.

How could one estimate what these weights might be? The easiest way to do this is to measure all of the variables on an entire cohort of children and then follow them for some time. A statistical analysis in which one predicts later academic accomplishments from the admission variables will show the relative importance of each variable in predicting the dependent variable. Within-ethnic-group analyses will show whether some variables are more or less important than others for different groups of children (see Tables 4 and 5 in Lohman, 2005, for one example). Longitudinal studies of this sort are one of the critical needs in the field of gifted education.

Identifying Academically Talented Minority Students

Prediction of Achievement for Minority Students

The conceptual and methodological guidelines discussed thus far generally apply to the identification of all academically talented students. But are modifications of these guidelines needed when identifying academically talented minority students? Do the same characteristics function as aptitudes? Concretely, are the predictors of academic achievement the same for majority and minority students? And even if they are the same, should they be weighted the same? For example, are nonverbal reasoning abilities more predictive of achievement for minority students than for majority students? Is the ability to reason with English words less predictive of achievement for Hispanic or Asian-American students than for White students?

We have examined this question in some detail. Our analyses, which concur with those of other investigators (e.g., Keith, 1999), are unequivocal: The predictors of achievement in Reading, Mathematics, Social Studies, and Science are the same for White, Black, Hispanic, and Asian-American students (Lohman, 2005).

For example, Figure 7 shows how scores on the three CogAT batteries combine to predict ITBS Reading Achievement. Two regression weights are shown above each path. These weights show the relative importance of each CogAT score for the prediction of reading achievement. The first weight is for non-Hispanic White students; the second (in parentheses) is for Hispanic students. Clearly, the predictors of success in reading are the same for both groups. The CogAT Verbal Battery is the strongest predictor; CogAT Nonverbal Battery contributes least to the prediction. Indeed, nonverbal reasoning abilities sometimes have a significant *negative* regression weight in the prediction of achievement once verbal and quantitative reasoning abilities are in the equation (Case, 1977; Lohman, 2005). This means that some students with high nonverbal reasoning scores are actually *less* likely to achieve in school than are other students with similar levels of verbal and quantitative abilities.

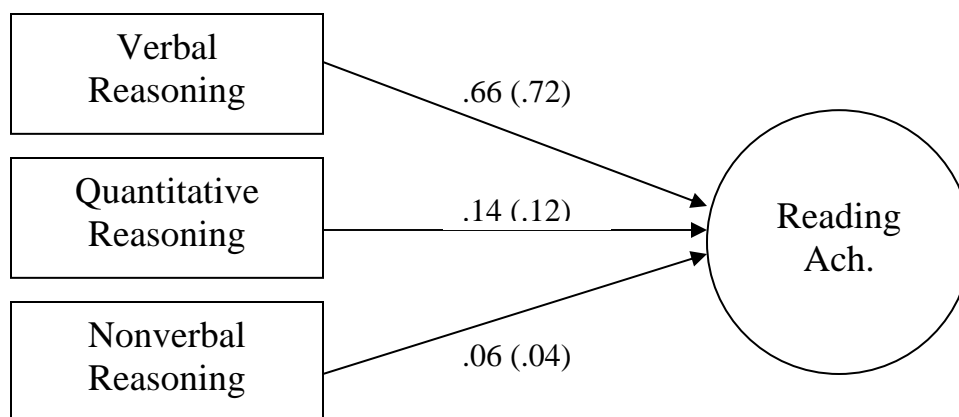


Figure 7. Relative contributions of CogAT Verbal, Quantitative, and Nonverbal reasoning abilities to the prediction of ITBS reading achievement for U.S. White and Hispanic students. The first value is the regression weight for non-Hispanic White students; the second (in parentheses) is for Hispanic students. The multiple correlations were $R = .81$ and $.80$ for White and Hispanic students, respectively.

This makes sense from the perspective of aptitude theory. Schooling places heavy demands on students' abilities to use language to express their thoughts and to understand other people's attempts to express their thoughts. Because of this, those students most likely to succeed in formal schooling in any culture will be those who are best able to reason verbally. Indeed, our data show that, if anything, verbal reasoning abilities are even *more* important for bilingual students than for monolingual students. This is because the student who does not have great familiarity with the English language frequently must infer the meanings of unfamiliar English words from contextual and other cues.

Thus, an aptitude perspective leads one to look for those students who have best developed the specific cognitive (and affective) aptitudes most required for acquiring expertise in particular domains. Identifying such students requires this attention to proximal, relevant aptitudes, not distal ones that have weaker psychological and statistical justification.

Assumptions About Growth

Predictions about future performance assume that a student's rank within the comparison group on the aptitude test will remain relatively constant over time. This does not mean one assumes that scores are fixed. Scores that report rank within age or grade group easily mask the fact that all abilities are developed; all respond to practice and instruction. Rather, the assumption is that a student's rate of growth on the skills measured by the test will be the same as other students in the norm group who obtained

the same initial score. This is unlikely either if the student's experiences to date differ from those of the norm group or if her subsequent experiences depart from the norm. For example, lack of experience in a domain will lead to a lower initial rank than will be achieved later as the student has the necessary learning experiences. This is especially true for well-defined skill sets that are quickly learned (e.g., learning the letters of the alphabet) rather than for open-ended skill sets that require extensive practice (e.g., verbal comprehension). However, a student can also fall behind over time by improving, but at a slower rate than her peers.

In general, prediction equations for academic success do not differ by ethnicity. Indeed, if anything, aptitude tests more often over-predict the academic performance of some minority students (Willingham, Lewis, Morgan, & Ramsit, 1990). Therefore, programs that aim to assist minority students in developing their academic talents might best understand their task as one of falsifying a prediction about growth rate.

This is not easily done. Contrary to popular myth, complex skills and deep conceptual knowledge do not suddenly emerge when the conditions that prevented or limited their growth are removed (cf. Humphreys, 1973). The attainment of academic excellence comes only after much practice and training. It requires the same level of commitment on the part of students, their families, and their schools as does the development of high levels of competence in athletics, music, or in other domains of nontrivial complexity. Further, because the relationship between aptitude and outcome is probabilistic, one cannot expect that every student who is identified as likely to succeed will do so. The critical issue for programs that aim to assist these children, however, is to maximize the proportion of identified students that do succeed.

Judging Test Bias by Mean Differences Rather Than by Predictive Validity

A selection policy that uses either ability or achievement tests alone or that combines, say, mathematics achievement and quantitative reasoning ability would select proportionately fewer Black and Hispanic students than White and Asian-American students. How, then, can one attend to the relevant aptitude variables and increase the representation of minority students served?

Note that the discussion in this section concerns the identification of high-aptitude—not high-achievement—students. Academic achievements and accomplishments, although perhaps measured in somewhat different ways for different individuals, should always be evaluated against the same high standards. That more White or Asian-American students achieve at high levels is problematic if the measurement procedures can be shown to be biased against other students. Measurement professionals agree that this is generally not the case (Jencks, 1998).

The identification of aptitude is a much slipperier task. Even in the best of circumstances, correlations between measures of aptitude and achievement at some future date are substantially less than perfect, so predictions will often be wrong. More importantly, *one can make inferences about aptitude from scores on a collection of tasks*

only when individuals have had similar opportunities to develop the skills required for success on those tasks. All recognize that many students—especially those whose first language is not English—have not had the same opportunities to develop skills in the English language. Therefore, many schools screen students with nonverbal tests, teacher questionnaires, and performance assessments because differences between ELL and native speakers of English are sometimes smaller on such tests.⁵

The need for a test that minimizes group differences is a consequence of the assumption that one must always compare every student to every other student in an age or grade cohort. There are many reasons for this. In part, it stems from the laudable desire to be fair. All children are compared to the same standards, or so it seems. In part, it stems from the failure to appreciate the extent to which the normative scores on ability tests shift monthly and on achievement tests shift weekly to accommodate slight differences in children's experiences in the culture or in school. And in part, it stems from the administrative convenience of using norms provided by the publisher rather than having to develop local or local subgroup norms. Other things being equal, we surely would prefer the assessment procedure that showed the smallest difference between ethnic groups. However, other things are rarely equal.

The consequences of assuming that test bias can be judged by differences in group means are generally overlooked. Some of the more obvious effects are that it

1. ***Reinforces the tendency to interpret intelligence and other ability tests as measuring innate abilities.*** If scores on ability tests depend on background and education, then one must take these factors into account when interpreting them. The alternative—to interpret test scores as measures of innate abilities largely unaffected by such factors—avoids these complications. Thus, the decision to use a common cut score on aptitude tests inadvertently encourages the naïve but false belief that ability tests measure innate rather than developed abilities.
2. ***Encourages the use of less reliable tests.*** The smaller the mean difference between groups on the selection test, the greater the proportion of students from lower-scoring groups who will be selected using a common cut score. In general, group differences will be smaller on less reliable tests than on more reliable tests. For example, performance tests are generally less reliable than objective tests and will generally show smaller group differences than objective tests. In the extreme, a completely unreliable test will show no differences between groups even when true differences are large. Therefore, *evaluating tests by the extent to which they achieve the goal of proportional representation will tend to favor shorter and otherwise less reliable tests over longer and more reliable tests.*
3. ***Encourages the use of old tests with outdated norms.*** More students (minority and majority) will attain high scores on a test with outdated norms than

⁵ Differences are especially large when nonverbal and verbal reasoning scores of ELL students are compared. Differences are much smaller between quantitative and nonverbal reasoning tests, especially for Asian-American students. As a group, Black students often perform better on verbal and quantitative tests than on nonverbal reasoning tests (see, e.g., Jencks & Phillips, 1998).

on a test with recent norms. If the test is administered to all students, then the proportion of minority students will be larger simply because the admission standard has been lowered (see point 4). If the test with older norms is only used to screen for minority students, then the proportion of minority students admitted will be even greater.

4. ***Encourages the lowering of standards.*** If the scores for two groups differ in their means and show equal variability, then lowering the admission standard will increase the proportion of students from the lower-scoring group. For example, assume one changes the cut from, say, the top 5% to the top 10% of cases. There will be a greater proportion of students from the lower-scoring group when the cut is set at the top 10% rather than at the top 5%. How much greater depends on the specific score distributions. Note, however, that the total number of students admitted has now doubled. A variant on this theme is to set different cut scores for different tests and then compare tests in terms of the proportion of minority students admitted. Other things being equal, the test with the lower cut score will admit proportionately more minority students.⁶

5. ***Encourages the use of less valid tests.*** The hope that one can use a common cut score for all applicants leads one to opt for selection tests on which group differences are smaller. In general, though, when differences in achievement are large, differences will also be large on measures that predict achievement. Tests that are less predictive of achievement are more likely to show somewhat smaller group differences. Using less valid tests and a common cut score, one may identify more minority students, but fewer who have the aptitude to succeed. This should be of concern to all, and especially to the minority communities who hope that the students who receive extra assistance will develop into the next generation of minority scholars and professionals.

The Need for Within-group Comparisons

A better policy, then, is to make decisions about *aptitude* for academic excellence using the most valid and reliable measures for all students, but to compare each student's scores only to the scores of other students who share roughly similar learning opportunities or background characteristics. In other words, inferences about aptitude should be made within such groups.

This does not mean that every child needs to be compared only to the handful of other children in the population who share her unique circumstances. Simply comparing

⁶ An easy way to envision this is to draw two normal distributions, with one positioned a bit to the right of the other. The proportion of students who obtain any score is given by the relative heights of the two distributions at that point on the score scale. For every point above the mean of the left distribution, the height of the right distribution will be greater. The further one moves to the right on the scale, the greater the discrepancy between the heights. At the extreme right, only students from the higher scoring distribution will be included. Put differently, small group differences at the mean translate into increasingly larger differences in the proportion of cases that form each group as one moves up the score scale.

an ELL student to all other ELL students in a grade will help enormously. If the sample is sufficiently large, then one can further subdivide into groups with little, intermediate, and extensive exposure to the English language (Ortiz & Ochoa, 2005). One of the most important advantages of group-administered ability tests is that they allow these sorts of comparisons when a common test is administered to all of the students in a school or district. Further, one need not derive formal norms to make such within-group comparisons. A simple rank-order of scores will often serve the purpose. This is demonstrated in the sample data set below.

A Sample Data Set

The identification policies advocated here are not difficult to implement. I have created a sample data set and instructions that detail the process. Those who would like to try the procedures can download the data set and instruction file from my website (http://faculty.education.uiowa.edu/dlohman/sample_data_set).

I created the data set by randomly sampling scores for 100 White, 100 Black, and 100 Hispanic third graders from a large data set that contains scores for students on both Form 6 of CogAT and Form A of the ITBS. To make the issues clear, I picked a data set that showed large differences between ethnic groups. Other than deleting cases with incomplete data, no changes were made before randomly sampling 100 students from each ethnic group. Here I will illustrate how to identify high-scoring minority students by combining CogAT and ITBS scores to use both for selection.

Table 1 shows the means and standard deviations of scores obtained by White, Black, and Hispanic students. For this sample, Whites scored highest on all three batteries of CogAT, Blacks obtained their highest score on the Verbal Battery, and Hispanics obtained their highest score on the Nonverbal Battery.

Table 1.

Means (and SDs) for CogAT6 SAS Scores, by Ethnicity for a Random Sample of 300 Students

	Ethnicity		
	White (<i>N</i> = 100)	Black (<i>N</i> = 100)	Hispanic (<i>N</i> = 100)
CogAT Battery			
Verbal Reasoning	102.4 (15.8)	93.0 (16.0)	90.0 (13.7)
Quantitative Reasoning	101.4 (15.8)	91.9 (15.7)	92.1 (14.1)
Nonverbal Reasoning	103.6 (15.9)	91.4 (15.2)	94.1 (15.4)

I have argued that the interpretation of scores involves comparisons to multiple norm groups: national, local, and opportunity-based subgroups. Here I used ethnicity as a surrogate for the opportunity-to-learn subgroup. A better procedure would be to form

experience-based subgroups. This is especially helpful for ESL students. In districts with large immigrant populations, one could easily group English as a Second Language (ESL) students in each grade into two or three subgroups on the basis of their exposure to and familiarity with the English language. In some districts, virtually all of these students will speak the same language; in other districts, many different languages will be represented. Although native language matters, experience with English matters more. In any case, the goal is to be able to see how the student's scores compare to all three of the relevant norm groups: national, local, and subgroup. The third norm group compares the student's scores to those of others who have had similar opportunities to learn the abilities sampled by the test.

There are several ways to do this. One of the easiest is to divide students into ethnic groups and then rank order their scores on the test. This works if there is only one test score to consider. However, for estimating academic aptitude for a domain one should always consider both achievement in the domain and the ability to reason in the symbol system(s) needed to acquire new knowledge in that domain. For elementary school children, the two most important achievement domains are reading and mathematics, and so the most important domains of reasoning are with verbal and quantitative symbols. In the verbal domain, the identification process should consider both current reading achievement and verbal reasoning ability.

How can one identify those students who, on average, score highest on both tests? One method is to plot of the pairs of scores for each student, making a separate scatter plot for students in each ethnic group. These scatter plots are shown in Figure 8. Each plot shows the relationship between individual students' CogAT Verbal Battery SAS scores and their ITBS Reading Total scores. Which students are most likely to show continued improvement in school? Those with the highest average score on *both* tests (see the CogAT + ITBS column in Table 2). These students are identified by solid circles. The plots show clearly those cases in which the two tests disagree. Students whose within-group rank is similar on both tests will fall near an imaginary line that runs through the middle of the swarm of points.

In this sample, the discrepancies between CogAT and ITBS scores are largest for the White students and smallest for the high-scoring Hispanic students. Large discrepancies should always be investigated. In some cases, these can be traced to problems the student might have experienced in taking one of the tests. Students represented by an "x" in Figure 8 responded inconsistently to the CogAT Verbal Battery (see Footnote 8).

Although scatter plots can help identify those students within each group who score high on both tests, a tabular presentation of the data has advantages. If the scores for all students are placed in a single table, one can see at a glance how each student's scores compare to all three norm groups: the nation, the local population, and those students who have had roughly similar opportunities to develop the skills that were tested. All three of these perspectives are useful.

Table 2 shows these data for 50 students with the highest scores on an equally weighted composite of the CogAT Verbal Battery and the ITBS Reading Total score. This composite score was formed by converting the CogAT Verbal SAS scores and ITBS Reading Total scores to standard or z -scores. This makes the SDs of the two scores the same. In Microsoft Excel, z -scores can be created by applying the "standardize" function. Then the two z -scores are simply summed. One can easily weight one variable more than the other by applying other weights to the z -scores (e.g., $2z_1 + 1z_2$). The cases are then sorted on this new composite score using the Microsoft Excel "sort" function.

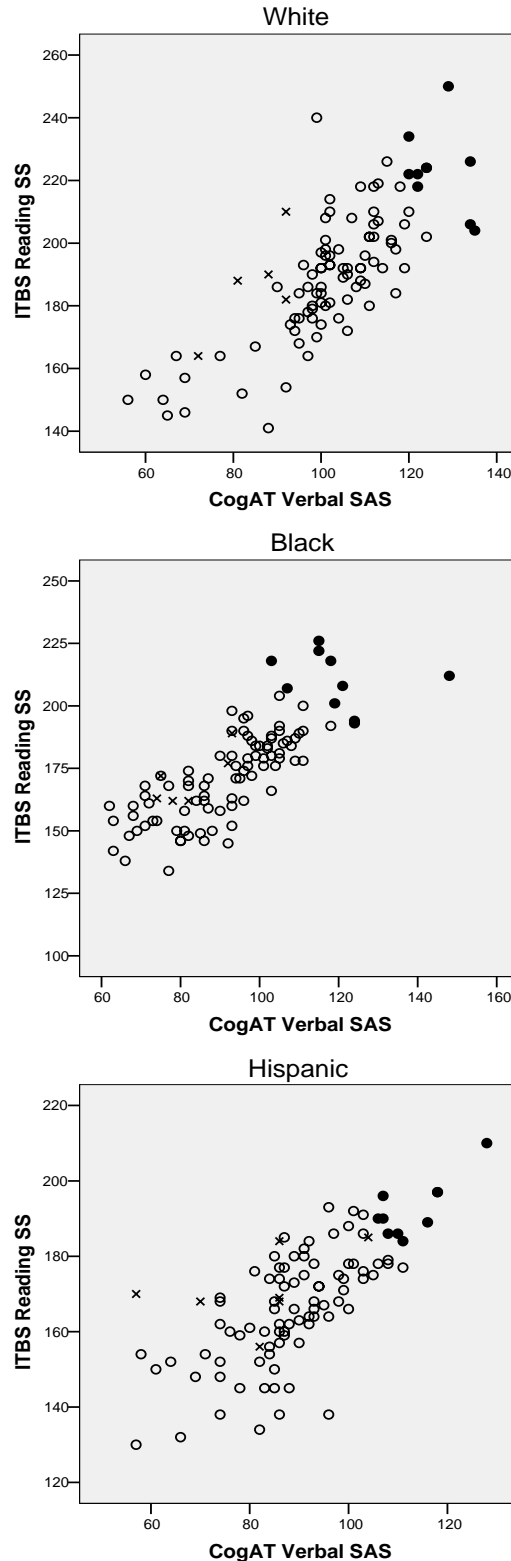


Figure 8. Plots of ITBS Reading versus CogAT Verbal Reasoning, by Ethnicity. Solid circles = top 10 in the group, x = flagged Verbal SAS score.

Table 2.

Students With the 50 Highest Scores on CogAT Verbal Battery + ITBS Reading Total

ID ¹	Gender	CogAT Verbal			CogAT Profile ²	ITBS Reading Total		CogAT+ ITBS	Local Rank
		V-SAS	V-PR	V-flag		Read SS	Read PR		
W 87	F	129	97	0	9A	250	99	5.49	1
B 92	M	148	99	0	9E (V+N-)	212	87	4.89	2
W 62	M	134	98	0	6E (V+)	226	96	4.67	3
W 79	F	120	89	0	7A	234	98	4.18	4
W 95	M	124	93	0	8A	224	95	3.96	5
W 99	M	124	93	0	8A	224	95	3.96	6
W 75	F	122	92	0	7C (V+Q-)	222	94	3.74	7
W 76	M	134	98	0	7E (V+)	206	81	3.74	8
W 70	M	135	99	0	8C (V+N-)	204	78	3.70	9
W100	M	120	89	0	8B (N+)	222	94	3.61	10
W 82	M	122	92	0	8A	218	92	3.55	11
H 99	M	128	96	0	8E (Q-)	210	85	3.55	12
W 55	M	115	83	0	6A	226	96	3.49	13
B100	F	115	83	0	8B (N+)	226	96	3.49	14
B 90	M	115	83	0	7A	222	94	3.30	15
W 89	F	118	87	0	7A	218	92	3.30	16
B 84	M	118	87	0	6B (V+)	218	92	3.30	17
W 8	M	99	48	0	4B (N-)	240	99	3.15	18
W 80	M	120	89	0	8C (Q+N-)	210	85	3.05	19
W 31	F	113	79	0	5B (V+)	219	92	3.04	20
B 71	M	121	91	0	7B (N-)	208	83	3.02	21
W 51	M	112	77	0	6A	218	92	2.93	22
W 90	F	124	93	0	8B (Q-)	202	76	2.93	23
W 94	F	119	88	0	7A	206	81	2.80	24
W 25	F	109	71	0	6B (N-)	218	92	2.74	25
B 97	M	119	88	0	8A	201	74	2.57	26
W 69	F	112	77	0	7A	210	85	2.55	27
B 94	M	124	93	0	7C (V+Q-)	194	64	2.55	28
B 89	F	124	93	0	6B (V+)	193	62	2.50	29
W 74	F	113	79	0	7B (Q-)	207	82	2.47	30
W 65	F	116	84	0	7A	201	74	2.38	31
B 99	F	103	57	0	7E (V-N+)	218	92	2.37	32
W 30	F	112	77	0	5E (V+Q-)	206	81	2.37	33
W 57	F	116	84	0	6A	200	73	2.33	34
H 82	F	118	87	0	6B (V+)	197	69	2.32	35
H 85	M	118	87	0	6C (V+Q-)	197	69	2.32	36
W 78	M	117	86	0	7A	198	70	2.30	37
W 56	F	112	77	0	7B (Q+)	202	76	2.18	38
W 93	M	107	67	0	6B (N+)	208	83	2.15	39
W 32	M	119	88	0	5B (V+)	192	61	2.14	40
W 77	F	102	55	0	7B (V-)	214	89	2.12	41
W 66	F	111	75	0	6B (Q+)	202	76	2.12	42
W 83	M	111	75	0	7A	202	76	2.12	43
B 78	M	107	67	0	6C (Q+N-)	207	82	2.10	44
B 70	M	118	87	0	5B (V+)	192	61	2.08	45
B 98	M	111	75	0	8B (V-)	200	73	2.02	46
W 91	F	102	55	0	8E (V-)	210	85	1.93	47
B 66	F	105	62	0	4B (V+)	204	78	1.84	48
W 37	M	114	81	0	5B (V+)	192	61	1.83	49
H 98	M	116	84	0	7C (Q-N+)	189	57	1.82	50

¹Ethnicity indicated by the first character in the ID.²Profile indicates the level and pattern of the CogAT scores. The number is the median age stanine. If the first letter is an A, all three battery scores were at approximately the same level; if a B, one battery score was above or below the other two; or if a C, there was a significant contrast between two battery scores with the third in between. An E profile indicates an extreme difference between the highest and lowest battery scores. Letters in parentheses explain the pattern (relative strengths and weaknesses).

The first character in the ID column of Table 2 indicates the student's ethnicity. To help identify subgroups, IDs for Black and Hispanic students have been highlighted. Notice that in addition to the CogAT Verbal Battery SAS, the table includes the national percentile rank (PR) for the Verbal SAS. Similar data are reported for the ITBS Reading standard score (SS) and its national percentile rank. All three norm groups can now be seen: The CogAT PR and ITBS PR scores show where the students rank nationally on each of the tests; their position in the sorted list shows where they rank locally, and their position within the coded subgroup shows where they rank within that subgroup of students.⁷

Other kinds of information in Table 2 include gender and a variable called "vflag." This variable reports whether the student's Verbal Battery score was flagged because the three CogAT verbal subtest scores were discrepant or because his responses to items were inconsistent.⁸ Scores for these students were indicated with an "x" in the scatterplots. If CogAT Verbal Battery scores are flagged for either of these reasons, the scores should not be used to make decisions about the student.

The variable labeled CogAT *profile* tells something about the student's scores on all three CogAT batteries. Notice that only 17 of the 50 students have an "A" score profile. An "A" profile means that the level of the Verbal, Quantitative, and Nonverbal Battery scores did not differ significantly. This is the profile that is assumed whenever one uses the CogAT Composite score (or a similar *g* measure) for selecting students. Clearly, most students are not well characterized by such a model. This is why the CogAT authors have long advised educators not to select students for TAG programs using the Composite score for all three batteries (Lohman & Hagen, 2001b, Thorndike & Hagen, 1978, 1987, 1993).

All four of the Hispanic students listed in Table 2 obtained higher PR scores on the CogAT Verbal Battery than on the ITBS Reading test. This is because the two tests make quite different demands on the students' English Language reading skills. By design, the ITBS accentuates differences in students' reading abilities, whereas CogAT attempts to minimize them. This does not mean that the Reading test is biased against Hispanic students. (Indeed, the Reading and Verbal reasoning tests are more closely related for these students than for White or Black students.) Rather, it means that, compared to the national norm group, these students show much greater ability to reason in the English language than to read it. However, the fact that they reason so well in

⁷ Note that a rank-ordered list does not provide the same sort of local percentile ranks that would be provided by a test publisher. The latter would reflect the proportion of the local population scoring lower than the examinee on a smoothed score distribution. The simpler ranks used here do not capture this information. But they are often all that is needed when the goal is to identify a certain percentage of the local or subgroup population.

⁸ Form 6 of CogAT is the only test that I know that cautions users if a student's responses are inconsistent on a test battery. This can be very helpful in understanding discrepancies between CogAT scores and scores on other tests. Note that all of the examinees whose CogAT Verbal scores were marked "x" in Figure 8 showed higher ITBS Reading scores. See the case study of the academically gifted student Maxwell in Lohman and Hagen (2001b, p. 90).

English means that, given intensive instruction, their reading abilities are likely to show rapid improvement.

Rank-ordered lists like those shown in Table 2 can be used to identify either those Black or Hispanic students who exceed a particular national PR, a particular local rank, or a particular subgroup rank. This portion of the table identifies the 13 Black and 4 Hispanic students with the highest scores on the combined CogAT Verbal Battery + ITBS Reading Total score. Those students who obtain the highest scores on this composite would be most similar to those served by the typical TAG program. Suppose the program serves the top 5% of students in the local population. In this population of 300 third graders, this would include students 1-15. This group would contain 11 White students, 3 Black students, and 1 Hispanic student. Many of these students would probably benefit from some kind of academic acceleration in verbal domain. Increasing the diversity of the students served by program, however, means identifying those Black and Hispanic students who have comparable ranks within their respective groups on this aptitude measure. The Black students who most clearly show strong verbal aptitude are those who come next on the list: students B-84, B-71, B-97, B-94, B-89, and B-99. For the Hispanic students, it is students H-82 and H-85. If the goal were to identify, say, the top 10 scorers in each ethnic group, then one would need to look further down the table for more Hispanic students. (These students are shown clearly in the scatter plot in Figure 8.) All have exhibited strong verbal aptitude when compared with others in the same ethnic group. However, the curricular needs of these students will generally not be the same as the curricular needs of students who scores placed them at the top of the overall list. Furthermore, when making decision about academic placements, evidence of achievement in the domain of instruction should take precedence over estimates of verbal reasoning or the composite verbal aptitude measure.

Table 3 shows a similar set of scores for quantitative reasoning abilities. Here, the summary variable is an equally weighted composite of z -scores for the CogAT Quantitative Battery SAS score and the ITBS Mathematics Total scaled score. This time the group of top 15 scorers contains 9 White students, 6 Black students, and 1 Hispanic student (who had the highest score of all students). There are also three more Hispanic students (H-93, H-98, and H-89) and three more Black students (B-91, B-71, and B-81) who have only slightly lower scores. This is not uncommon. Minority students often show excellent mathematics achievement. Indeed, quantitative reasoning abilities are often an important strength of Black high-school students (Lohman, 2004). Depending on the curriculum, students with slightly lower scores high-accomplishment group on the composite measure of mathematics aptitude might profit from the same instructional arrangements as students with higher scores. However, there is no mechanical way to make this judgment. Much depends on the demands of the mathematics curriculum and other characteristics of these students, especially their motivation.

Table 3.

Students With the 50 Highest Scores on CogAT Quantitative Battery + ITBS Math Total

ID ¹	sex	CogAT Quantitative			CogAT Profile ²	ITBS Mathematics Total		CogAT+ ITBS	Local Rank
		Q-SAS	Q-PR	Q-flag		Math Total SS	Math PR		
H 94	M	147	99	0	7E (V-Q+)	223	99	5.628	1
W 95	M	136	99	0	8A	226	99	5.101	2
W 87	F	130	97	0	9A	226	99	4.722	3
W 82	M	124	93	0	8A	225	99	4.287	4
W 69	F	121	91	0	7A	219	97	3.759	5
W100	M	121	91	0	8B (N+)	219	97	3.759	6
B 90	M	117	86	0	7A	223	99	3.732	7
W 91	F	127	95	0	8E (V-)	210	91	3.632	8
W 92	M	135	99	0	8E (V-)	201	80	3.631	9
B 98	M	129	97	0	8B (V-)	207	88	3.590	10
W 99	M	125	94	0	8A	211	92	3.562	11
B 94	M	110	73	0	7C (V+Q-)	226	99	3.458	12
W 96	F	126	95	0	8B (V-)	208	89	3.456	13
B 92	M	130	97	0	9E (V+N-)	202	81	3.372	14
B 95	M	130	97	0	7B (Q+)	202	81	3.372	15
W 56	F	124	93	0	7B (Q+)	206	86	3.217	16
W 79	F	118	87	0	7A	212	93	3.176	17
H 93	F	124	93	0	7A	205	85	3.161	18
W 70	M	121	91	0	8C (V+N-)	206	86	3.028	19
H 98	M	109	71	0	7C (Q-N+)	219	97	3.001	20
H 89	M	132	98	0	7B (Q+)	193	66	2.992	21
B 91	F	130	97	0	6E (Q+)	195	70	2.978	22
W 65	F	121	91	0	7A	205	85	2.972	23
W 83	M	112	77	0	7A	215	95	2.966	24
B 71	M	116	84	0	7B (N-)	210	91	2.937	25
W 66	F	125	94	0	6B (Q+)	199	76	2.887	26
B 81	M	114	81	0	6B (N-)	211	92	2.867	27
W 25	F	109	71	0	6B (N-)	215	95	2.776	28
W 59	F	116	84	0	6B (V-)	205	85	2.656	29
W 55	M	108	69	0	6A	213	94	2.600	30
W 90	F	112	77	0	8B (Q-)	207	88	2.515	31
W 80	M	127	95	0	8C (Q+N-)	190	59	2.507	32
W 84	M	113	79	0	7B (V-)	205	85	2.466	33
H 70	M	108	69	0	5C (V-Q+)	210	91	2.431	34
W 81	M	124	93	0	7E (V-)	192	64	2.430	35
H 74	F	117	86	0	6B (Q+)	199	76	2.381	36
W 62	M	107	67	0	6E (V+)	210	91	2.368	37
H 83	M	115	83	0	6C (V-Q+)	201	80	2.367	38
H 82	F	106	65	0	6B (V+)	211	92	2.361	39
W 22	F	113	79	0	5B (Q+)	203	82	2.353	40
W 77	F	121	91	0	7B (V-)	194	68	2.353	41
W 85	M	113	79	0	7A	202	81	2.297	42
B100	F	120	89	0	8B (N+)	194	68	2.289	43
W 93	M	111	75	0	6B (N+)	203	82	2.227	44
W 46	M	101	52	0	5A	214	94	2.214	45
W 89	F	114	81	0	7A	197	73	2.079	46
B 97	M	120	89	0	8A	190	59	2.064	47
B 84	M	102	55	0	6B (V+)	210	91	2.052	48
W 94	F	113	79	0	7A	197	73	2.016	49
B 80	M	109	71	0	6A	201	80	1.988	50

¹ Ethnicity indicated by the first character in the ID.

² Profile indicates the level and pattern of the CogAT scores. The number is the median age stanine. If the first letter is an A, all three battery scores were at approximately the same level; if a B, one battery score was above or below the other two; or if a C, there was a significant Contrast between two scores with the third in between. An E profile indicates an Extreme difference between the highest and lowest battery scores. Letters in parentheses explain the pattern (relative strengths and weaknesses).

Overall, then, the process of identifying the most academically promising minority students is the same as the process of identifying the most academically promising majority students. First, identify those aptitudes that research shows best predict academic success. Next, measure the ability, achievement, interest, and other variables for all students. Try to get multiple measures of each and average them. For example, average students' reading achievement scores obtained on the same test across two years. Third, before combining scores for different constructs, first put them on the same scale. This is easily done in any spreadsheet. Fourth, sort the data by this new composite measure. Fifth, identify students using primarily their within-group rank on the composite measure, but consider other factors as well. Sixth, make decisions about academic placement using rank compared to all other students in the local population. Seventh, reassess at regular intervals. Expect new students to qualify for special services and some students who have needed these services in the past not to need them. Giftedness is not a state of being, but a statement about the current rank of a student on an ever-changing scale cognitive development and academic accomplishment.

Caveat: Selection as an Ill-structured Problem

Aptitude for any complex endeavor has many components. Estimating academic aptitude requires considering cognitive abilities as well as current academic accomplishments, motivation, interest, willingness to work with others, and other factors that moderate success in the particular types of instructional programs that are (or can be) offered. Identification is always an ill-structured problem for which there is no one best solution. Therefore, it is generally helpful to have more rather than less information at hand. However, it is also important to know how to integrate this information to make good decisions. Checklists or matrices can provide useful ways to organize the variables, but they cannot tell how best to combine them. Some factors deserve much weight; others deserve less weight or can even be ignored at times. The empirical evidence clearly supports giving primary weight to evidence of current accomplishments and reasoning abilities in those symbol systems needed to create new understandings in the domain. Although affective factors are important, the weight given to particular measures depends on the kind of instructional program that the student will face. For example, some children will thrive if paired with a mentor with whom they identify. For these children, the social dimension of learning is critical. Other children enjoy learning about the domain itself and will learn much even if they have access only to texts or a computer. Therefore, one must consider many factors when making decisions about which children to admit to a program or, alternatively, which kind of instructional arrangement might best fit the needs of a particular student.

Although adapting instruction to the needs of students is a critical aspect of any successful program, I have not emphasized it here. Too often, children are labeled "gifted" on the basis of an IQ test; other affective and cognitive aptitudes required for success are ignored or are only considered after the student has been identified. This has it backwards. Giftedness means superior aptitude or talent for something, not for everything. Programs would do a better job of identifying talented children if they started with a clear understanding of the types of expertise that they seek to develop and

the kinds of instruction that they can offer. Together, these will more clearly define the personal characteristics that will function as aptitudes for success in those programs.

Paradoxically, this approach is even more important for identifying high-aptitude minority students than for identifying students who already display exceptional academic accomplishment. For example, fear of rejection by peers is pervasive among many under-achieving minority students. If the student does not value academic achievement then success in any program is unlikely until that critical aspect of readiness has been developed or the program has been modified to accommodate its absence. Therefore, programs that aim to assist talented students who do not share the worldview of middle-class America must look beyond the measurement of cognitive competence. However, it is impossible to adapt a program better to meet the needs of particular group of students until one knows clearly the source of the mismatch between those students and the demands of the program. *Therefore, the aptitude approach described in this monograph applies not only to the identification of those students most likely to succeed in a given program. It also is a critical step in making effective modifications of programs better to serve the needs of these students.*

Suggestions for Policy

How can educators implement a policy consistent with the principles outlined here?

1. *What are the purposes of the TAG program?* Is the emphasis on *T* (Talent) or *G* (Gifted)? Is the goal to identify and serve those students who demonstrate unusually high levels of academic ability and accomplishment? If so, then traditional procedures of identifying and serving academically "gifted" students can be used. Poor and minority students will be included in this group, although not at a level that approaches their representation in the population. Attempts to achieve greater minority representation by using nonverbal tests and other measures that are not good measures of scholastic aptitude will indeed include more ELL students in the program. Unfortunately, these will not in general be the most academically promising students. On the other hand, if the goal is to identify the most academically talented students in underrepresented populations regardless of current levels of academic attainment, then procedures like those outlined in this paper will be more successful. However, options for educational placement and programming will need to be much more diverse than is currently the case. Perhaps in this way, TAG programs could infuse procedures for identifying academic talent and then providing developmentally appropriate instruction into mainstream educational practices. It is not only academically gifted students who are not well served by a rigidly age-tracked educational system.

2. *Identify better, more psychologically defensible methods for identifying the most academically talented minority students.* Discuss the difference between the need for common national and local standards for the measurement of current achievement and the need for within-group standards for the measurement of aptitude. Setting common, high standards for all encourages those who do not yet display academic skills to work toward them. Because both estimates of aptitude and accomplishment will be lowest for

those who have had the fewest opportunities, consider grouping students by opportunity to learn and making identification within groups. Then make instructional placements primarily on the basis of accomplishments to date. Keep in mind that there is also an ethical dimension to consider. For some children, the intensive instruction offered in special programs for the talented provides opportunities that supplement what their families provide; for other children, the same programs provide the *only* opportunity to develop such skills. Indeed, the goal for these students is to provide educational opportunities that will falsify the prediction that, as a group, future achievement will show the same or lower rank than current achievement.

3. *What educational treatment options are available?* Understanding the programs that are or can be offered by the school is the first step in identifying which personal characteristics will function as aptitudes (or inaptitudes) for those programs. In what content areas can advanced instruction be offered? Will students receive accelerated instruction with age-mates? Or will they attend class with older children whose achievement is at approximately the same level? Will instruction require much independent learning, or must the student work with other students? Will instruction build on students' interests, or is the curriculum decided in advance? Are mentors available who can encourage and work with those students who need extra assistance? These different instructional arrangements will require somewhat different cognitive, affective, and conative aptitudes. At the very least, different instructional paths should be available for those who already exhibit high accomplishment and for those who display talent but somewhat lower accomplishment. For all students, acceleration of one sort or another is often the least expensive way to provide developmentally appropriate instruction (Colangelo, Assouline, & Gross, 2004). If schools cannot provide this sort of differential placement, then it is unlikely that they will be able to satisfy the twin goals of providing developmentally appropriate instruction for academically advanced students while simultaneously increasing the number of underrepresented minority students who are served and who subsequently develop academic excellence.

4. *Obtain the most reliable and valid measures of achievement, reasoning abilities, and other aptitude variables for all students.* Whenever possible, measure the behavior of interest rather than something that merely predicts that behavior. For example, if interested in children's weight, then weigh them. Do not measure their heights and try to predict weight from height. Similarly, if the goal is to identify students who have unusual talent for particular academic domains, obtain measures of domain-specific achievement, the student's ability to reason in the symbol systems required for new learning in that field of study, interest in the domain, and persistence under similar instructional conditions. For example, to identify students who excel in mathematics, first measure mathematics achievement using a well-constructed, norm-referenced achievement test that emphasizes problem solving and concepts. To identify students who are most likely to show the strongest future development, combine scores on the mathematics achievement test with scores on measures of quantitative and figural reasoning abilities. Combine the scores in a way that weighs mathematics achievement and reasoning abilities equally. To assess interests, inquire specifically about the students' interests in mathematics or in occupations that require mathematical thinking.

Interest inventories can be helpful, especially for adolescents (see Lubinski et al., 1995). Finally, estimate persistence, anxiety, and other important affective aptitudes from ratings obtained from teachers and others who have worked with the child in situations similar to those in the planned acceleration program. Keep in mind that aptitude can only be estimated when a student's performance on a task is compared with the performance of other students who have had similar learning opportunities. Common cut scores on less valid and reliable tests may identify significant numbers of minority students, but many of them are not the students who have the greatest academic talent.

5. Make better use of local norms on both ability and achievement tests, especially when identifying students whose accomplishments in particular academic domains are well above those of their classmates. On norm-referenced tests, examine local percentile ranks for particular domains such as mathematics or science rather than national percentile ranks for composite scores. This will facilitate the goal of providing challenging instruction for all students. When making instructional placements, use local norms to determine the appropriateness of the match. For example, if a student will be placed with seventh graders for mathematics, compare her performance on a test with seventh grade mathematics content to the performance of students in the prospective seventh grade class.

6. Emphasize that true academic giftedness is evidenced by accomplishment. Predictions that one might someday exhibit excellence in a domain are flattering but unhelpful if they do not translate into purposeful striving toward the goal of academic excellence. Indeed, the attainment of academic excellence requires the same level of commitment on the part of students, their families, and their schools as does the development of high levels of competence in any other domain. Students may find it helpful to consider selection for special academic programming as analogous to being identified as a "high-potential" athlete, and then discuss the duration and intensity of training that high-caliber athletes endure to rise to the top of their sport. This also means that students must be identified with an eye on the kind of intensive instruction that can be offered. If advanced instruction will be in writing short stories, then measures of quantitative or figural reasoning abilities will not identify many of those who are most likely to succeed. Further, if possible, the instruction that is offered should be adapted better to meet the needs of minority students. On the affective side, eliciting interest and persistence are critical. On the cognitive side, oral language skills in the dialect of the language students are expected to read and write are probably the most neglected, but among the most important aptitudes for academic success. Many suggestions can be derived from case studies of successful minority scholars or from evaluations of schools that routinely produce them (e.g., Presseley, Raphael, Gallagher, & DiBella, 2004).

7. Professional judgment is required. There is no foolproof way to identify those children who will develop the highest levels of academic excellence in adolescence or the highest levels of professional expertise as adults. Simple schemes that establish an arbitrary cut score on an IQ or achievement test are administratively convenient but identify only a fraction of those who will later attain excellence. Further, such schemes necessarily disadvantage children who have had fewer opportunities to develop the

abilities measured by the tests at the time selections are made. One can go a long way toward correcting this bias by identifying the most academically talented students within the opportunity-to-learn groups. This does not mean that new measures are not needed or would not be helpful. My reading of the research, however, says that we would probably do better by looking for new ways to measure domain knowledge, interest, and motivation than by continuing to search for better measures of general reasoning and problem-solving abilities. The most pernicious assumption is that somehow, someday, one will find a way to measure true (i.e., innate) ability that will be independent of culture, education, opportunity to learn, and motivation. This is not possible. However, it is possible that we will find new and better ways to measure those aptitudes that are required for later success that can usefully supplement the measures currently in hand. And some of these new measures may show smaller ethnic group differences than existing ability and achievement tests (as seems to be the case for measures of verbal creativity and fluency). However, the primary evidence needed to support the use of such measures in identification is their ability to contribute to the long-term prediction of academic success. This requires longitudinal studies that investigate relationships between aptitude measures and subsequent measures of accomplishment both for all students as well as for groups of students who have had substantially different opportunities to acquire the knowledge, skills, and other attributes measured by the aptitude tests. Importantly, we already have well-documented ways for identifying academically promising students that are as effective for minority students as for majority students. We just need to learn how to use these methods more intelligently.

References

- Ackerman, P. L. (2003). Aptitude complexes and trait complexes. *Educational Psychologist*, 38, 85-94.
- Anastasi, A. (1937). *Differential psychology*. New York: Macmillan.
- Anastasi, A. (1980). Abilities and the measurement of achievement. In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade: Proceedings of the 1979 ETS Invitational Conference* (pp. 1-10). San Francisco: Jossey Bass. (*New directions for testing and measurement series*)
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York: Harper & Brother.
- Boothe, D., & Stanley, J. C. (Eds.). (2004). *In the eyes of the beholder: Critical issues for diversity in gifted education*. Waco, TX: Prufrock Press.
- Bracken, B. A., & McCallum, R. A. (1998). *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside.
- Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Mind, brain, experience, and school*. Washington, DC: National Academy Press.
- Carroll, J. B. (1974). The aptitude-achievement distinction: The case of foreign language aptitude and proficiency. In D. R. Green (Ed.), *The aptitude-achievement distinction* (pp. 286-303). Monterey, CA: CTB/McGraw-Hill.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, England: Cambridge University Press.
- Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: A tribute to Arthur Jensen* (pp. 5-21). Oxford, England: Elsevier.
- Case, M. E. (1977). *A validation study of the Nonverbal Battery of the Cognitive Abilities Test at grades 3, 4, and 6*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
- Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, 40, 153-193.

- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54, 1-22.
- Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton-Mifflin.
- Cattell, R. B., & Cattell, K. S. (1965). *Manual for the Culture-Fair Intelligence Test, Scale 2*. Champaign, IL: Institute for Personality and Ability Testing.
- Colangelo, N., Assouline, S. G., & Gross, M. U. M. (2004). *A nation deceived: How schools hold back America's brightest students* (Vol. 1 & 2). Iowa City, IA: The University of Iowa, The Connie Belin & Jacqueline N. Black International Center for Gifted Education and Talent Development.
- Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., & Talbert, J. E. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.
- Cronbach, L. J. (1976). Measured mental abilities: Lingered questions and loose ends. In B. D. Davis & P. Flaherty (Eds.), *Human diversity: Its causes and social significance* (pp. 207-222). Cambridge, MA: Ballinger.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.
- Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.
- Degler, C. (1991). *In search of human nature: The decline and revival of Darwinism in American social thought*. New York: Oxford University Press.
- Donovan, M. S., & Cross, C. T. (Eds.). (2002). *Minority students in special and gifted education*. Washington, DC: National Academy Press.
- Ferguson, G. A. (1956). On transfer and the abilities of man. *Canadian Journal of Psychology*, 10, 121-131.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Roco, Peterson.
- Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, 101, 171-191.

- Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.
- Gardner, H. (2003, April). *Multiple intelligences after twenty years*. Paper presented at the meeting of the American Educational Research Association, Chicago. Retrieved August 11, 2005, from Harvard University, Project Zero Website: <http://www.pz.harvard.edu/PIs/HG.htm>
- Glaser, R. (1992). Expert knowledge and processes of thinking. In D. F. Halpern (Ed.), *Enhancing thinking skills in the sciences and mathematics* (pp. 63-75). Hillsdale, NJ: Lawrence Erlbaum.
- Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-46). New York: Macmillan.
- Gustafsson, J. -E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 4, pp. 35-71). Hillsdale, NJ: Lawrence Erlbaum.
- Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). *Iowa Test of Basic Skills: Form A*. Itasca, IL: Riverside.
- Horgan, J. (1995). Get smart, take a test. *Scientific American*, 273(5), pp. 12, 14.
- Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53-91). New York: Guilford Press.
- Horn, J. L., & Blankson, N. (2005). Foundations for better understanding cognitive abilities. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 41-68). New York: Guilford Press.
- Humphreys, L. G. (1973). Implications of group differences for test interpretation. In *Proceedings of the 1972 Invitational Conference on Testing Problems: Assessment in a Pluralistic Society* (pp. 56-71). Princeton, NJ: ETS.
- Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87-102). New York: Plenum.
- Humphreys, L. G., & Davey, T. C. (1988). Continuity in intellectual growth from 12 months to 9 years. *Intelligence*, 12, 183-197.
- Hunt, E. (2000). Let's hear it for crystallized intelligence. *Learning and Individual Differences*, 12, 123-130.

- Hunt, E. B., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 7, pp. 87-122). New York: Academic Press.
- Hunt, J. M. (1961). *Intelligence and experience*. New York: Ronald Press.
- Jencks, C. (1998). Racial bias in testing. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 55-85). Washington, DC: Brookings Institution Press.
- Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White test score gap*. Washington, DC: Brookings Institution Press.
- Jensen, A. R. (1981). Raising the IQ: The Ramey and Haskins study. *Intelligence*, 5, 21-40.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Kelley, T. L. (1927). *Interpretation of educational measurements*. Chicago: World Book.
- 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.
- Lawson, A. E. (2004). Reasoning and brain function. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 12-48). Cambridge, England: Cambridge University Press.
- Lewis, J. D. (2001). Language isn't needed: Nonverbal assessments and gifted learners. *Proceedings of the Growing Partnerships for Rural Special Education Conference*. San Diego, CA. (ERIC Document Reproduction Service No. ED 453026)
- Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology*, 3, 482-494.
- Lohman, D. F. (1993). Teaching and testing to develop fluid abilities. *Educational Researcher*, 22, 12-23.
- Lohman, D. F. (1994). Spatial ability. In R. J. Sternberg (Ed.), *Encyclopedia of Intelligence* (pp. 1000-1007). New York: Macmillan.
- Lohman, D. F. (2004). Aptitude for college: The importance of reasoning tests for minority admissions. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in college admissions* (pp. 41-56). New York: Routledge/Falmer.

- Lohman, D. F. (2005). The role of nonverbal ability tests in the identification of academically gifted students: An aptitude perspective. *Gifted Child Quarterly*, 49, 111-138.
- Lohman, D. F. (in press-a). An aptitude perspective on talent: Implications for the identification of academically gifted minority students. *Journal for the Education of the Gifted*.
- Lohman, D. F. (in press-b). Beliefs about differences between ability and accomplishment: From folk theories to cognitive science. *Roeper Review*.
- Lohman, D. F., & Hagen, E. P. (2001a). *Cognitive Abilities Test (Form 6)*. Itasca, IL: Riverside.
- Lohman, D. F., & Hagen, E. P. (2001b). *Cognitive Abilities Test (Form 6): Interpretive guide for school administrators*. Itasca, IL: Riverside.
- Lohman, D. F., & Korb, K. A. (in press). Gifted today but not tomorrow? Longitudinal changes in *ITBS* and *CogAT* scores during elementary school. *Journal for the Education of the Gifted*.
- Lorge, I., Thorndike, R. L., & Hagen, E. (1964). *The Lorge-Thorndike Intelligence Tests*. New York: Houghton Mifflin.
- Lubinski, D., & Benbow, C. P. (2000). States of excellence. *American Psychologist*, 55, 137-150.
- Lubinski, D., Benbow, C. P., & Ryan, J. (1995). The stability of vocational interests among the intellectually gifted from adolescence to adulthood: A 15-year longitudinal study. *Journal of Applied Psychology*, 80, 196-200.
- Marland, S. P. (1972). *Education of the gifted and talented. Vol. 1. Report to the Congress of the United States by the U.S. Commissioner of Education*. Washington, DC: U.S. Government Printing Office.
- McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 136-181). New York: Guilford Press.
- Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test: Multilevel technical manual*. San Antonio, TX: Harcourt Brace.
- 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., & Ford, D. Y. (2005). Increasing minority children's participation in gifted classes using the NNAT: A response to Lohman. *Gifted Child Quarterly*, 49, 29-36.
- Nelson, C. A. (1999). Neural plasticity and human development. *Current Directions in Psychological Science*, 8, 42-45.
- Nickerson, R. S. (1998). Confirmation bias: Ubiquitous phenomenon in many guises. *Review of General Psychology*, 2, 410-433.
- Nickerson, R. S. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Lawrence Erlbaum.
- Ortiz, S. O., & Ochoa, S. H. (2005). Advances in cognitive assessment of culturally and linguistically diverse individuals. In D. P. Flanagan & P. L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, Tests, and Issues* (2nd ed., pp. 234-250). New York: Guilford Press.
- Otis, A. S., & Lennon, R. T. (1997). *Otis-Lennon School Ability Test* (7th ed.). San Antonio, TX: Harcourt Brace.
- Peterson, P. L. (1977). Interactive effects of student anxiety, achievement orientation, and teacher behavior on student achievement and attitude. *Journal of Educational Psychology*, 69, 779-792.
- Presseley, M., Raphael, L., Gallagher, J. D., & DiBella, J. (2004). Providence-St. Mel School: How a school that works for African American students works. *Journal of Educational Psychology*, 96, 216-235.
- Raven, J. C. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, 41, 1-48.
- Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's Progressive Matrices and vocabulary scales, section 4: Advanced Progressive Matrices, sets I and II*. London: H. K. Lewis.
- Renzulli, J. S., Smith, L. H., White, A. J., Callahan, C. M., Hartman, R. K., & Westberg, K. L. (2002). *Scales for Rating the Behavioral Characteristics of Superior Students*. Mansfield Center, CT: Creative Learning Press.
- Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, 40, 163-184.
- Roid, G. (2003). *Stanford Binet Intelligence Scales, technical manual* (5th ed.). Itasca, IL: Riverside.

- Ruf, D. L. (2003). *Use of the SB5 in the assessment of high abilities* (Stanford-Binet Intelligence Scales, Fifth Edition, Assessment Service Bulletin No. 3). Itasca, IL: Riverside.
- Scarr, S. (1994). Culture-fair and culture-free tests. In R. J. Sternberg (Ed.), *Encyclopedia of Human Intelligence* (pp. 322–328). New York: Macmillan.
- Schmidt, D. B., Lubinski, D., & Benbow, C. P. (1998). Validity of assessing educational-vocational preference dimensions among intellectually talented 13-year-olds. *Journal of Counseling Psychology, 45*, 436-453.
- 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.
- Snow, R. E. (1978). Theory and method for research on aptitude processes. *Intelligence, 2*, 225-278.
- Snow, R. E. (1980). Aptitude and achievement. In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade: Proceedings of the 1979 ETS Invitational Conference* (pp. 39-60). San Francisco: Jossey Bass. (*New directions for testing and measurement series*)
- Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist, 27*, 5-32.
- Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context: Interactionist perspectives on human intelligence* (pp. 3-37). Cambridge, England: Cambridge University Press.
- Snow, R. E., & Lohman, D. F. (1984). Toward a theory of aptitude for learning from instruction. *Journal of Educational Psychology, 76*, 347-376.
- Snow, R. E., & Yalow, E. (1982). Education and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 493-585). Cambridge, England: Cambridge University Press.
- Spearman, C. E. (1923). *The nature of intelligence and the principles of cognition*. London: Macmillan.

- 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.
- Stern, W. (1914). The psychological methods of testing intelligence (*Educational Psychology Monographs No. 13*). Baltimore: Warwick & York. (G. M. Whipple, trans.)
- Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Lawrence Erlbaum.
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.
- Subotnik, R., & Jarvin, L. (2005). Beyond expertise: Conceptions of giftedness as great performance. In R. J. Sternberg & J. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 343-57). New York: Cambridge University Press.
- Sundet, J. M., Tambs, K., Magnus, P., & Berg, K. (1988). On the question of secular trends in the heritability of intelligence test scores. *Intelligence*, 12, 47-59.
- Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. New York: Columbia University, Teachers College.
- Thorndike, R. L., & Hagen, E. (1978). *Cognitive Abilities Test (Form 3)*. New York: Houghton Mifflin.
- Thorndike, R. L., & Hagen, E. (1987). *Cognitive Abilities Test (Form 4)*. Chicago: Riverside.
- Thorndike, R. L., & Hagen, E. (1993). *Cognitive Abilities Test (Form 5)*. Chicago: Riverside.
- Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science*, 14, 623-628.
- VanTassel-Baska, J. (2000). The on-going dilemma of effective identification practices in gifted education. *The Communicator*, 31, 39-41.
- Willingham, W. W., Lewis, C., Morgan, R., & Ramsit, L. (1990). *Predicting college grades: An analysis of institutional trends over two decades*. New York: The College Board.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside.