The Role of Nonverbal Ability Tests in Identifying Academically Gifted Students: An Aptitude Perspective

David F. Lohman
The University of Iowa

ABSTRACT

The first goal of this article is to discuss the role of nonverbal ability tests in the identification of academically gifted children. I note that most nonverbal tests measure verbally mediated cognitive processes, that they are neither "culture free" nor "culture fair," and that we have known these facts for a very long time. I show that selecting students for gifted and talented programs on the basis of such tests would exclude the majority of the most academically accomplished students in all ethnic groups. The second goal is to propose a better method for identifying gifted students. I argue that the critical issue is readiness for a particular type of educational opportunity. The cognitive aspects of readiness are evidenced first in students' levels of knowledge and skill in particular domains and secondarily in their abilities to reason in the symbol systems used to communicate new knowledge in these domains. This applies to both minority and majority students. Therefore, the most academically talented minority students are those who show the strongest current achievement in particular domains and the best ability to reason in the symbol systems required for the acquisition of new knowledge in those domains. I also argue that, although current accomplishment can be measured on a common scale, judgments about potential must always be made relative to circumstances.

I first learned about nonverbal ability tests in the early 1970s when I was taught how to administer many of these tests to hearing-impaired students at the school for the deaf where I was working. By the mid-1970s I was in graduate school working on a research project that aimed to understand the cognitive processes people used when attempting to solve items on ability tests of all sorts.

PUTTING THE RESEARCH TO USE

Discovering which characteristics to measure on selection tests requires a careful consideration of the knowledge, skills, motivation, and other personal attributes that are required for success in particular academic programs. At the very least, programs for the gifted need to distinguish between the academic needs of students who currently show academic excellence and the needs of those who show lesser accomplishments, but have potential for developing academic excellence. The most important aptitudes for future academic accomplishment in a domain are current achievement in that domain and the ability to reason in the symbol systems in which new knowledge is communicated. For both minority and non-minority students, verbal and quantitative reasoning abilities are much better predictors of academic success than nonverbal, figural reasoning abilities. In fact, some students with high nonverbal abilities are actually *less* likely than other students to develop academic excellence. Further, many of the most academically capable Black students score poorly on such tests. Although accomplishments can be estimated using common norms, potential must always be judged relative to circumstances. It is recommended, therefore, that programs use common aptitude measures, but uncommon cutoff scores (e.g., rank within group) when identifying those minority students most likely to profit from intensive instruction. Tests of nonverbal, figural reasoning abilities are a helpful adjunct for both minority and nonminority admissions—but evidence shows that they should be measures of last resort, not first resort. When used alone, such tests increase selection bias while appearing to reduce it.

Because figural tests are particularly amenable to such inquiry, much of our work centered on these tasks (Snow & Lohman, 1984, 1989). Over the years, I conducted many studies on spatial abilities, figural reasoning abilities, and the nature of individual differences in thinking, problem solving, and their implications for instruction. In the early 1990s, I was asked if I would assume responsibility for the sixth edition of the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001a).

As one who has spent 30 years studying figural reasoning and as the coauthor of an ability test that has an excellent nonverbal battery, one might expect that I would be pleased with the recent emphasis on using figural reasoning tests to identify students for inclusion in programs for the gifted and talented. On the contrary, I am dismayed by the claims that have been made for such tests. I see well-intentioned educators being misled about what these tests measure and, more importantly, children being hurt by selection policies that use nonverbal reasoning tests as the criteria of first resort—rather than of last resort—for admission to programs for the academically gifted and talented.

The goals for using figural reasoning tests when selecting students for special programs for the gifted and talented are laudable: Measure abilities in a way that is fair to all students; increase the diversity of students who are included in programs for the gifted and talented; actively assist those who have not had the advantages of wealth or an immersion from birth in the English language. I endorse these goals. I also believe that figural reasoning tests can provide information that assists in achieving them. Such tests have a place at the selection table. But, I disagree with those who claim that they should be at the head of the table or, worse yet, occupy the only chair at the table.

Contributors to this journal have disagreed on the role of nonverbal ability tests in the identification of academically gifted students. For example, Naglieri and Ford (2003) advocated the use of Naglieri's (1997) group-administered figural reasoning test for identifying academically gifted students. However, in a previous article (Lohman, 2005), I showed that the score distributions for different ethnic groups used in their study had been substantially altered from those reported in previous analyses of the same data. The claims about the utility of the test for identifying gifted minority students were therefore not supported. Mills and Tissot (1995) also counseled caution. They noted that large differences in the mean scores of ethnic groups and low correlations between scores on the Advanced Progressive Matrices

Test (APM; Raven, Court, & Raven, 1983) and measures of achievement make it a poor primary selection tool for special programs that involve advanced coursework. As they put it: "Identification instruments should match the programs for which students are being identified" (p. 216). In their view, a more appropriate use of the APM may be as a screening test that, along with other assessments, could be used to identify academic potential in students who are not yet ready for advanced-level academic programs. Such students could be provided educational opportunities that aim to develop academic skills needed to participate in advanced-level coursework. I concur with their conclusions, although I also would argue that measures of quantitative and verbal reasoning should generally be considered before the nonverbal-reasoning test in the identification process. Like Richert (2003), I also argue that rank within group on the most relevant aptitudes should guide efforts to identify academically promising minority students who are not yet ready for advanced-level academic programs.

Overview

I first give a brief overview of different types of nonverbal ability tests. These include a broad range of group and individual tests that measure an equally broad range of abilities. A major point is that 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 examinees use to solve items. I then argue that claims that such tests are "culture fair" mislead because they encourage the mistaken belief that abilities can be measured in ways that are independent of culture, experience, and motivation. This is not possible.

I then focus on nonverbal reasoning tests (e.g., the Raven Progressive Matrices) that require examinees to reason with figural stimuli. I argue that such tests should not be the primary instrument for identifying academically gifted students. I first show that selecting students on the basis of such tests would exclude most of the students who would profit from advanced instruction and include many who would not profit from it.

Understanding why this is so requires a grasp of the nature of human abilities, how they develop, and how they function as aptitudes for future learning. I begin by noting the correspondence between physical and mental abilities. This helps make clear the claim that all abilities

are developed and that all tests measure developed abilities. I then discuss why it is important to distinguish among the abilities to reason with verbal, quantitative, and spatial concepts. I conclude this section by showing that a relative strength in spatial abilities seems to be an inaptitude for aspects of academic learning.

I next turn to the larger question of how we might best identify those students who either presently or some time in the future would most profit from advanced instruction. I argue that the primary question to be addressed in selecting students for special programs is one of readiness for a particular type of educational opportunity, not innate ability. Readiness has cognitive, affective, and conative dimensions. I show that the cognitive aspects of readiness are evidenced first and foremost in students' levels of knowledge and skill in particular domains and secondarily by their abilities to reason in the symbol systems used to communicate new knowledge in those domains. Figural reasoning tests are generally distal predictors of readiness for academic learning. Importantly, the predictors of current and future academic excellence are the same for minority and majority students. This means that the most academically talented minority students are those who show the strongest current achievement in particular domains and the best ability to reason in the symbol systems required for the acquisition of new knowledge in those domains.

Finally, I argue that programs for the academically gifted should distinguish between high levels of current accomplishment in a domain and lesser levels of current accomplishment, but potential for higher levels of future accomplishment. Acceleration to more advanced classes or instruction at an advanced level is often warranted for the high-accomplishment group, whereas intensive instruction somewhat above that received by age peers is often more appropriate for the high-potential group. Because judgments about potential are much more probabilistic than judgments about accomplishment, fixed cutoff scores for identifying high-potential students are difficult to defend, especially when students come from markedly different backgrounds.

Two caveats at the outset: First, giftedness has many manifestations. Here I discuss only academic giftedness. This is not a judgment about the importance of musical or athletic or other types of giftedness, but a necessary concession to the limitations of space. Second, although identification of academically gifted students should include many sources of information (Assouline, 2003; Hagen, 1980), I focus on the role of ability tests—especially nonverbal ability tests—in the process.

Nonverbal Ability Tests

General Characteristics of Nonverbal Tests

Tests are commonly called *nonverbal* if items present visual stimuli such as concrete objects or line drawings and require a nonverbal response such as assembling a puzzle, pointing to an answer, or filling in a circle under a picture. Directions may be given verbally, in pantomime, or through feedback on the correctness of responses given to a set of practice items. Verbal directions are more common on group-administered tests, whereas pantomimed directions or modeling of the desired behavior are more common on individually administered tests.

Verbal Processes in Nonverbal Tests. To call a test *nonverbal* is to make a statement about the test stimuli, not the cognitive processes examinees use to solve test items.[1] Indeed, "nonverbal items" commonly either require verbal or mathematical knowledge or use tasks whose solution is greatly facilitated by the use of verbal or mathematical cognitive processes. In some cases, these requirements are explicit. In some, they are less obvious.

An example of the explicit involvement of verbal processes is provided by the Pictorial Categories subtest of the Comprehensive Test of Nonverbal Intelligence (Hammill, Pearson, & Wiederholt, 1996). Each item on this subtest shows pictures of two objects at the top of the page and an empty box beneath them. The examinee must point to the object at the bottom of the page that goes in the box. The two pictures may be of an apple and a banana. The correct answer is another fruit rather than, say, a vegetable. A similar item in a verbal format would present the names of the objects, rather than line drawings of them. Verbal analogies can also be presented in pictures, as on the Analogic Reasoning subtest of the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998). For example, some items present line drawings of analogies such as "pear is to apple as carrot is to (a) grapes, (b) squash, (c) tomato, or (d) radish." The major difference between the verbal and nonverbal format for such items is that the examinee is expected to comprehend spoken or written words in the verbal test and to decipher the line drawings in the nonverbal test.

The pictorial format has both advantages and disadvantages. On the one hand, although the child must know words for both the objects depicted and the categories to which each belongs, that knowledge can be in any language. This is one of the main reasons why such tests are helpful when testing students with limited proficiency in

English, especially when proficiency is so limited that the student cannot understand orally presented test directions. On the other hand, it can be difficult for the examinee to decipher a line drawing (e.g., Is it an egg or a lemon?). Further, there are important regional and cultural differences in the familiarity of different objects and conventions in how they are depicted. For example, in the example above, there are regional and cultural differences in the vegetables children are most likely to see used at home. Particular items can also be harder or easier for students who speak different languages because of unexpected associations among words used to label the pictures.

On other tests, particularly those that use figural stimuli such as geometric shapes, the involvement of verbal processes is less obvious, but equally important. The Figure Analogies subtest on the CogAT and the matrix completion items on the Raven Progressive Matrices and the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997) are good examples of these types of tasks. Careful analyses of how examinees solve items on these tests (e.g., Bethell-Fox, Lohman, & Snow, 1984; Carpenter, Just, & Shell, 1990; Mulholland, Pellegrino, & Glaser, 1980; Sternberg, 1982) show that the labeling of stimuli (e.g., as a square or diamond), of the attributes of that stimulus (e.g., large, shaded with vertical lines), and of the rule that will generate the missing elements (the figures get bigger and the background is combined) are critical for successful solution of all but the simplest items. Failure to label figures, their attributes, or the transformations that are performed on them substantially increases the working memory burden on the examinee. Since the largest source of individual differences on reasoning tests is working memory burden (Kyllonen & Christal, 1990), anything that increases this burden can significantly impair test performance. Indeed, the major source of error on figural reasoning tasks is forgetting a transformation, attribute, or element (Lohman, 2000).

Understanding What to Do. The brief directions that are used on many nonverbal tests can create problems. For example, the most common mistake children make on analogy items is to pick a response alternative that is merely an associate of the last term in the stem (Achenbach, 1970; Sternberg & Nigro, 1980). They are more likely to do this if they do not really understand how to solve analogies. Cryptic or pantomime directions do not instruct them in the process of saying to themselves, "Pear is to apple as carrot is to what?," and then checking that the relationship they have inferred between the first two terms can indeed be mapped onto the second pair of terms. The problem of not really understand-

ing the directions goes beyond the analogy format. One study of the Raven Matrices showed that many minority children did not understand the directions. Going through the directions twice dramatically improved the scores of many of these students (Scarr, 1981). Therefore, eliminating written or spoken words does not somehow render problems the same for all, nor does reducing or eliminating verbal directions somehow level the playing field. In fact, it more commonly raises the slope for some.[2]

Confounding Reasoning and Spatial Abilities. Finally, some figural tests measure spatial abilities either inadvertently or explicitly. Most good tests of spatial ability require the examinee to perform an analog transformation on a mental image, such as mentally turning or rotating it to a new position. Therefore, figural reasoning tests that require these sorts of processes are particularly likely to measure spatial abilities, as well as reasoning abilities. Unless one intends to measure spatial abilities, this is not a good thing. The presence of sex differences provides a good way to distinguish between figural tests that measure spatial ability and those that measure reasoning abilities with figural stimuli. Good tests of spatial ability will show effect sizes for sex of .5 *SD* or more, whereas good nonverbal reasoning tests show no sex differences.[3]

Nonverbal Ability Tests as Culture–Fair Measures of g?

Some test authors are once again claiming that their nonverbal reasoning tests are "culture fair." The "culture-fair" claim is a less extreme version of an earlier claim that such tests were "culture free." However, the intuitively plausible notion that nonverbal reasoning tests are "culture free" or "culture fair" has been roundly criticized by both measurement specialists and cognitive psychologists. The notion of a "culture-free" test surfaced in the 1920s, became popular in the 1930s, but was debunked by some of the more thoughtful measurement experts in the 1940s. Cattell (1971) tried to resurrect the concept in the exposition of his theory of fluid versus crystallized abilities. Cattell attempted to avoid some of the more blatantly false assumptions of a "culture-free" test by calling it "culture fair." While many psychologists eventually became convinced of the utility of his concepts of fluid and crystallized abilities, the notion of "culture-fair" tests continued to be widely criticized (Anastasi & Urbina, 1997; Cronbach, 1990; Scarr, 1994).

The belief that one can measure reasoning ability in a way that eliminates the effects of culture is a recurring

fallacy in measurement. Culture permeates nearly all interactions with the environment; indeed, the concept of intelligence is itself rooted in culture (Sternberg, 1985). Further, nonverbal tests such as the Raven Progressive Matrices do not measure the same functions as verbal tests (Scarr, 1994), often show larger differences between ethnic groups than verbal or quantitative tests (Jensen, 1998), and are particularly liable to practice and training (Irving, 1983). Indeed, as Scarr (1994) noted, "Although tests such as the Raven Matrices may seem fair because they sample skills that are learned by nearly everyone . . . puzzle-like tests turn out to have their own limitations" (p. 324).

At the surface level, the claim that a test is "culture fair" means that the stimulus materials are assumed to be equally familiar to individuals from different cultures. Although there are cultures in which stylized pictorial stimuli are novel (Miller, 1997), children who have lived in developed countries are generally all exposed to common geometric shapes and line drawings of some sort. However, they may not be equally familiar with the names of these objects or as practiced in using those names. Stylized pictures of everyday objects often differ across cultures and within cultures across time.[4] Thus, the assumption that the test stimuli are equally familiar to all is dubious (Laboratory of Comparative Human Cognition, 1982, p. 687).

At a deeper level, though, the claim is that the types of cognitive tasks posed by the items—and thus the cognitive processes children must use when solving them—are equally familiar. There is an aspect of problem solving that is clearly rooted in culture, namely the habit of translating events into words and talking about them. Although children may recognize ovals, triangles, and trapezoids and may know about making things bigger or shading them with horizontal rather than vertical lines, the habit of labeling and talking aloud about such things varies across cultures (Heath, 1983).[5] Children who do not actively label objects and transformations are more likely to resort to a purely perceptual strategy on nonverbal tests. Such strategies often succeed on the easiest items that require the completion of a visual pattern or a perceptually salient series, but fail on more difficult items that require the identification and application of multiple transformations on multiple stimuli (Carpenter, Just, & Shell, 1990).

Thus, although less extreme than the "culture-free" claim, the "culture-fair" claim is equally misleading. Both claims help perpetuate the myth that "real" abilities are innate; that culture, experience, and education are contaminants; and that intelligence is a unidimensional, rather than a multidimensional concept. We have long known that, as Anastasi and Urbina (1997) observed, the very concept of intelligence is rooted in culture. Modern theories of intelligence begin with this fact (Sternberg, 1985). Importantly, they do not end there. Most go on to try to identify those cognitive structures and processes that generate observed differences on tasks valued as indicants of intelligence. But, experience *always* moderates these interactions, and formal schooling organizes tasks that provide opportunities for these experiences. Because of this, intelligence becomes, as Snow and Yalow (1982) put it, "education's most important product, as well as its most important raw material" (p. 496). Indeed, education actively aims to cultivate intelligence (Martinez, 2000). Educators who work with children who learn quickly and deeply from school have the most to lose from the misconception that intelligence is independent of experience. If abilities developed independently of experience, then what need would we have for enrichment or acceleration or, indeed, for education at all? The myth that very able children will do fine if left to their own devices is rooted in this misconception.

The Prediction Efficiencies of Figural Reasoning Tests

Figural reasoning tests, then, are one important variety of nonverbal ability tests. Examples include the Raven Progressive Matrices (Raven et al., 1983), the Naglieri Nonverbal Ability Test (Naglieri, 1997), and the Figure Analogies, Figure Classification, and Figure Analysis subtests of the Cognitive Abilities Test (Lohman & Hagen, 2001a). These sorts of tests are sometimes used when screening students for inclusion in programs for the gifted, and strong claims have been made for their usefulness in making such decisions. Therefore, I focus exclusively on these sorts of tests in the remainder of this article.

The first claim that I make is that these sorts of nonverbal figural reasoning tests should not be the primary selection instruments for programs for the academically gifted and talented. The reasons typically given for their use are (a) scores on such tests show correlations with academic achievement that, while lower than the correlations between verbal or quantitative reasoning tests and achievement, are certainly substantial; and (b) differences between some (but not all) minority groups and English-speaking White students are smaller on figural reasoning tests than on tests with verbal content. Reduced mean differences make a common cutoff score seem more

acceptable when identifying children for inclusion in programs. Many also erroneously assume that the nonverbal test is a culture-fair measure of ability.

The reasons such tests should not be used as the primary selection tool are equally straightforward. Students who most need advanced academic instruction are those who currently display academic excellence. Although reasoning abilities are important aptitudes for academic learning, they are not good measures of current academic accomplishment. Further, of the three major reasoning abilities, figural reasoning ability is the most distal aptitude for success in the primary domains of academic learning (e.g., achievement in literacy or language arts, reading, writing, mathematics, science, and social studies). Selecting students for gifted and talented programs on the basis of a test of nonverbal reasoning ability would admit many students who are unprepared for—and thus would not profit from—advanced instruction in literacy, language arts, mathematics, science, or other content-rich domains. *It would also not select, and thereby exclude, many students who either have already demonstrated high levels of accomplishment in one of these domains or whose high verbal or quantitative reasoning abilities make them much more likely to succeed in such programs.* It would be like selecting athletes for advanced training in basketball or swimming or ballet on the basis of their running speed. These abilities are correlated, and running is even one of the requisite skills in basketball, but it is not the fair or proper way to make such decisions. Further, the teams selected in this way would not only include a large number of athletes unprepared for the training that was offered, but would exclude many who would actually benefit from it. Rather, the best measure of the ability to swim or play basketball or perform ballet is a direct measure of the ability to swim or play basketball or perform ballet. In other words, the primary measure of academic giftedness is not something that predicts academic accomplishment, but direct evidence of academic accomplishment (Hagen, 1980).

Understanding why a test that shows what some would consider a "strong" correlation with achievement should not be used as a substitute for the measure of achievement requires knowledge of how to interpret correlations. Sadly, many people who must rely on tests to make selection decisions do not understand how imprecise the predictions are, even from seemingly large correlations. Figure 1 shows an example of what a scatterplot looks like for a correlation of *r* = .6, which is a reasonable estimate of the correlation between a nonverbal ability test and a concurrently administered mathematics achievement test for both minority and nonminority students

(Naglieri & Ronning, 2000). Here, the nonverbal ability test is on the *x*-axis and a measure of mathematics achievement is on the *y*-axis. The percentile-rank (PR) scale is used since this is the common metric in selection. Suppose that we used the nonverbal reasoning test to identify students for a gifted and talented program and that we admitted the top 5%. How many students with mathematics achievement scores in the top 5% would be identified? In this particular sample, picking the top 5% on the nonverbal reasoning test would identify only four students who also scored in the top 5% on the mathematics achievement test. Two students actually scored below the sample median (PR = 50) on the achievement test. In general, picking the top 5% on the ability test would identify only 31% of the students in the top 5% of the math achievement test. Put differently, it would exclude 69% of the students with the best mathematics achievement. Further, about 10% of those who were selected would actually have scored below the mean on the mathematics test. Someday, these students may be ready for advanced instruction in mathematics, but clearly they have less need for it than the 69% of students with very high math scores who would be excluded. The situation is even worse if fewer students are selected (e.g., top 3%) or if the criterion is a verbal competency (such as writing) that has an even lower correlation with performance on the nonverbal test.

But, is it not true that nonverbal reasoning tests are good measures of *g*? Those who study the organization of human abilities using factor analyses routinely find that nonverbal reasoning tests are good measures of fluid reasoning ability (Gustafsson & Undheim, 1996). However, such analyses look only at that portion of the variation in test scores that is *shared* with other tests that are included in the factor analysis. Variation that is specific to the test is discarded from the analysis. Those who use test scores for selection get both parts, not just that portion of the shared variation that measures *g*. Unfortunately, the specific variance on figural reasoning tests is typically as large as the variation that is explained by the *g* or *Gf* factor.[6] Furthermore, the skills that are specific to the figural test are only rarely required in formal schooling. Indeed, as I will later show, some of these spatial skills may actually interfere with academic learning. This is not true for verbal or quantitative reasoning tests, in which most of the specific verbal or quantitative abilities measured are also required for success in school. Therefore, if we are interested in identifying those students most in need of acceleration in mathematics, social studies, or literature, then a reasoning test—especially a figural reasoning test—should
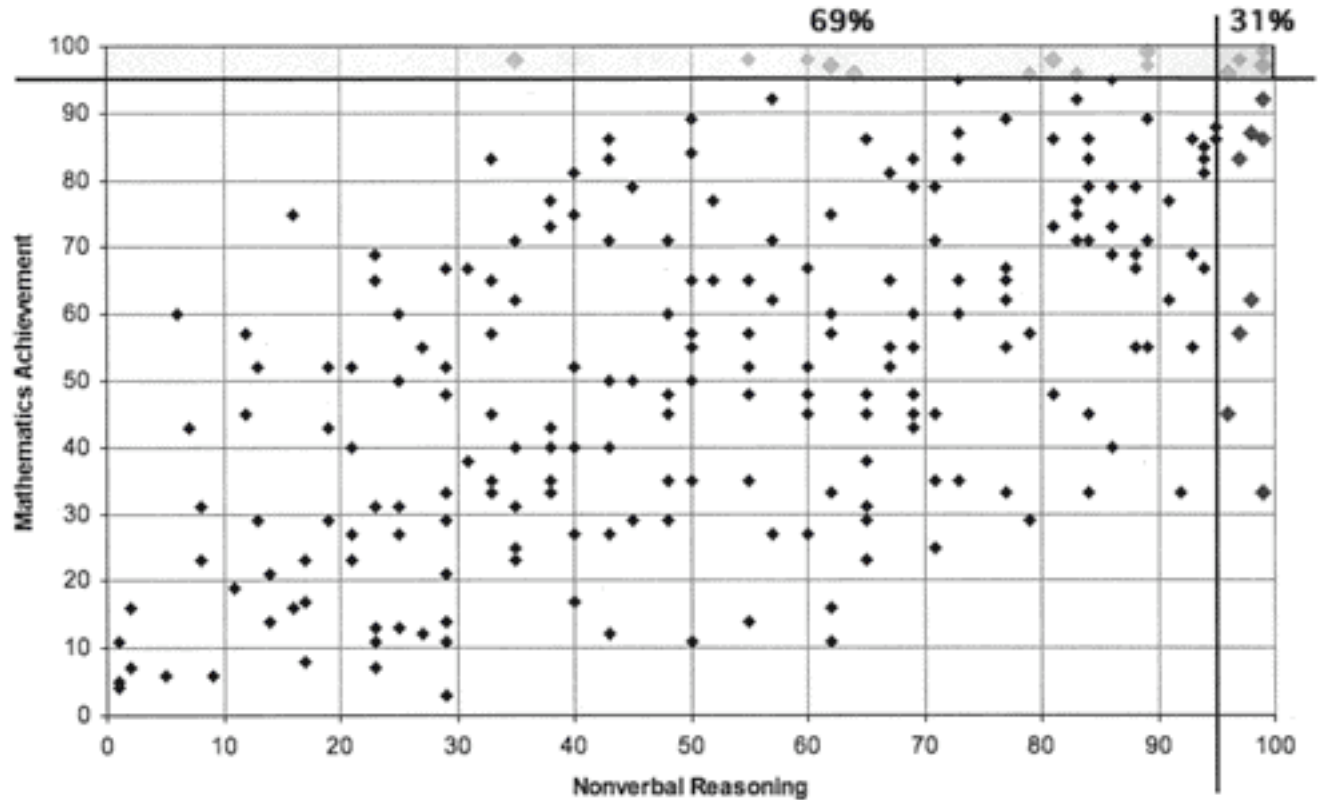
Figure 1. Example of a correlation of r = .6 between a nonverbal ability test (abscissa) and a mathematics achievement test (ordinate).

not be the primary selection instrument. Later, I will also show that a nonverbal reasoning test is also not the best way to identify those students who are most likely to *develop* high levels of achievement in academic domains. I would not want to be the program coordinator saddled with the responsibility of explaining the fairness of such a test to the parents of the many extremely high-achieving, but excluded students. I would also not want to be the administrator saddled with the responsibility of defending such a procedure in court.

Predicting Achievement for ELL Students

Would a figural reasoning test be more appropriate for identifying gifted English language learners (ELL) who perform well on tests that use a language other than English? Naglieri and Ronning (2000) reported correlations between the NNAT and Apprenda 2, an achievement test written in Spanish. The mean correlation between the NNAT and Spanish-language reading was *r* = .32. This means that picking Hispanic students for a program for gifted and talented students on the basis of their NNAT scores would generally exclude 80% of

those who read well in Spanish (i.e., score at or above the 90th percentile on the Apprenda 2). Figural reasoning abilities are not the same as verbal reasoning abilities in any language.

Distinguishing Achievement From Reasoning Abilities

Some think that it is unfair to use language-based tests of any sort to estimate abilities of bilingual or multi-lingual students. In some cases, this is because they want a test that measures the full extent of a child's verbal competence. This is understandable when identification is based on rank in the total sample, rather than rank within the subgroup of students with similar linguistic experience.

Others argue that ability and achievement are, as the words imply, distinct constructs. Like intuitive or folk theories in other domains, such theories are difficult to change even in the face of overwhelming contradictory evidence. The theory goes hand in hand with the belief that ability tests should be culture-free measures of innate capacity. Ability is treated like a mental tool that can be

applied to different intellectual content. Psychologically, the theory assumes that cognitive processes can be separated from knowledge. Although many psychologists hold such beliefs early in their research careers, eventually most discover that, not only is process enmeshed in knowledge, but much that appears to be process is actually the disguised effect of knowledge. For example, reasoning is better understood as a fancy name for search and comparison processes that are performed on a rich knowledge base. The sophistication of the reasoning that can display depends very much on what one knows and how that knowledge is represented in the brain.

Rather than separate circles for ability and achievement, most scholars envision a single universe of human competencies or, as Anastasi (1980) calls them, "developed abilities." In this universe, context is critical for both the development of ability and the expression of it. For example, words (and other symbol systems) not only express thought, but give birth to new ways of thinking. One cannot measure the sophistication of a child's reasoning or problem-solving abilities without embedding the problem in a context that elicits what the child knows and can do. To do so results in serious underrepresentation of the construct one hopes to measure. Braden (2000) noted that advocates of nonverbal testing recognize that language-based tests may introduce construct-irrelevant variance into the testing session for some students. However, they seem less aware that restricting test content results in construct underrepresentation—which is the other primary threat to test score validity. In other words, for children who speak or have learned some basic mathematical concepts, not using words or quantitative concepts prohibits them from showing how well they can reason. This is as true for minority students as it is for nonminority students

A test of reasoning abilities should therefore not seek to measure reasoning in contexts divorced from the conceptual knowledge in which it is grounded. Rather, the goal is, like a gardener, to prune the overgrowth and clear the weeds so that the flower stands out clearly.

Concretely, items on good verbal reasoning tests are constructed to emphasize reasoning processes and to reduce as much as possible the influence of extraneous factors such as word frequency. Consider, for example, the verbal analogies subtest of the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001a) that is administered to 12th graders. The typical correct answer is a word that can be used correctly in a sentence by about 75% of seventh graders. The average vocabulary level of all other words in the analogy items is grade 5. Nevertheless, the

analogy items are quite difficult. The typical 12th-grade student answers only about half of the items correctly. Indeed, well-constructed vocabulary tests that use relatively common, but abstract words are among the best measures of verbal reasoning. Students learn most new words by inferring plausible meanings from the contexts in which the words are embedded and then remembering and revising these hypotheses as they encounter the words anew. Achieving precise understandings of relatively common, but abstract words is thus an excellent measure of the efficacy of past reasoning processes in many hundreds or thousands of contexts. On the other hand, knowledge of infrequent or specialized words, while sometimes useful as a measure of prior achievement, estimates reasoning poorly.

Inferring missing words in a sentence is another ancient measure of verbal reasoning. However, Naglieri and Ford (2005) were particularly critical of the CogAT Sentence Completion subtest. They looked at the distribution of readability scores for 20 items at on level using the Flesch-Kinkaid method. The scores for sentences varied widely; the average grade-level score was 6.1 (a value they called "alarming"). However, their data have no merit. Readability formulas should never be used unless the passage has at least 100 words (Oakland & Lane, 2004). Even then they are of dubious value. Using readability formulas on sentences essentially produces a random variable. If reading items were a problem, then the difficulties of items should be predicted by their readability. However, the two sets of numbers are not significantly correlated. In other words, there is no evidence that students miss Sentence Completion items because they cannot read them.

Quantitative reasoning tests are particularly useful for identifying minority and ELL students who are likely to benefit from acceleration. The verbal requirements of such tests are minimal. Indeed, the directions are often shorter than directions for unfamiliar figural reasoning tests. Unlike figural reasoning, quantitative reasoning is an aptitude for a specific type of educational expertise that is developed in schools and thus affords enrichment and acceleration. Further, minority and ELL students generally perform quite well on such tests—often better than on tests of figural reasoning abilities. Finally, some argue that quantitative reasoning is actually a better marker for g than figural reasoning (Keith & Witta, 1997).

Performance on all ability tests reflects the complex interaction of biological preparedness and experience. Indeed, performance on figural reasoning tests (such as the Progressive Matrices Test and adaptations of it) is

markedly affected by education and practice. The so-called Flynn effect is much larger for such tests than for more educationally loaded tests (Flynn, 1987, 1999). Further, careful studies show the heritability of scores on such tests to be the same as the heritability of scores on achievement tests.[7] In other words, figural reasoning tests do not measure something that is any more (or less) the product of experience than good verbal reasoning tests.

Understanding Abilities

The Correspondence Between Physical and Mental Abilities

Although the relative influence of biology and experience varies across tasks, ultimately all abilities are developed through experience and exercise. However, the development of abilities is difficult to see because our intuitive theories of intelligence constantly get in the way. These intuitive theories are difficult to change because we cannot directly observe thinking or its development. If we could, we would see that cognitive and physical skills develop in much the same way. Because of this, it is helpful to consider the development of observable physical skills. Indeed, from Galton (1869/1972), to Bartlett (1932), to Piaget (1952), to cognitive psychologists such as Anderson (1982), theories of cognitive skills have been built on theories of physical skills. Anderson is most explicit about this; his model of the acquisition of cognitive skills is taken directly from Fitts' (1964) model of the acquisition of physical skills.

The correspondence between physical and cognitive abilities is shown graphically in Figure 2. Tests of general fluid abilities are akin to measures of general physical fitness. Measures of crystallized achievements in mathematics or literature, for example, are like observed proficiencies in particular sports such as basketball or swimming. Physical fitness is an aptitude for learning different sports. Those individuals with high levels of fitness generally find it easier to learn physically demanding activities and to do better at these activities once they learn them. In like manner, reasoning abilities are aptitudes for learning cognitively challenging subjects. Those who reason well learn more quickly and perform at higher levels once they have learned. Skilled athletic performance requires both biological preparedness and extensive practice and training. This is also true of complex cognitive skills.

However, physical fitness is also an outcome of participation in physically demanding activities. In like man-

ner, students who learn how to prove theorems in a geometry class or evaluate source documents in a history class also learn how to reason in more sophisticated ways. Thus, reasoning abilities are critical aptitudes for learning difficult material, as well as important outcomes of such learning.

Arguing that a good measure of reasoning ability should be independent of motivation, experience, education, or culture is like saying that a good measure of physical fitness should somehow be independent of every sport or physical activity in which the person has engaged. Such a measure is impossible. All abilities—physical and cognitive—are developed through exercise and experience. There are no exceptions.

Note that the analogy to physical skills provides an important role for biology. Speed, strength, and aerobic capacity are clearly rooted in inherited biological structures and processes. The analogy also suggests the importance of multiple test formats in the estimation of abilities. No test gives a pure estimate of ability. Tests that use the same format for all test items offer an advantage for students who (for whatever reason) do well on that format. This is particularly important for nonverbal reasoning tests because task specificity is generally much larger for figural tests than for verbal or quantitative tests (Lohman, 1996). Using a single-item format is like estimating physical fitness from a series of running competitions, rather than from a more varied set of physical activities.

The analogy to physical skills can also clarify why good measures of aptitude for specific academic domains such as mathematics or rhetoric must go beyond measures of figural reasoning ability. Success in ballet requires a somewhat different set of physical skills and propensities than success in swimming or basketball. A common set of running competitions would not be the best or fairest way to select athletes for advanced training in any of these domains, even if we could assume that all students had equal opportunities to practice running.

The Triad of Reasoning Abilities

There is now overwhelming evidence that human abilities are multidimensional, not unidimensional. This does not mean that, as Gardner (1983) would have it, g is unnecessary or unimportant (see Lubinski & Benbow, 1995). At the other extreme, it does not mean that g is the only thing that matters. Instead, it means that one must attend to both the overall level and the pattern of those abilities that are most important for school learning. This

is particularly important when attempting to identify gifted children.

The importance of going beyond *g* to measure a profile of reasoning abilities for all students (minority and majority) is shown clearly in the CogAT standardization data. Understanding why this is the case requires a brief review of how reasoning abilities are represented in hierarchical theories of human abilities. Carroll's (1993) three-stratum theory posits a large array of specific, or Stratum I, abilities (Carroll identified 69). These narrow abilities may be grouped into eight broad, or Stratum II, abilities. Stratum II abilities in turn define a general (*g*) cognitive ability factor at the third level. Importantly, the broad abilities at Stratum II vary in their proximity to the *g* factor at Stratum III. The closest is the broad fluid reasoning, or *Gf* factor.

Carroll's (1993) analyses of the fluid reasoning factor show that it, in turn, is defined by three reasoning abilities: (a) *sequential reasoning* (verbal, logical, or deductive reasoning); (b) *quantitative reasoning* (inductive or deductive reasoning with quantitative concepts); and (c) *inductive reasoning* (typically measured with figural tasks). These correspond roughly with the three CogAT batteries: verbal reasoning, quantitative reasoning, and figural/nonverbal reasoning. Each of these three reasoning abilities is estimated from two tests in grades K–2 and from three tests in grades 3–12.[8]

If given 90 minutes to test students' abilities, most psychologists would not administer a battery of nine different reasoning tests. Instead, they would try to represent a much broader slice of the Stratum II or Stratum III abilities in Carroll's model. Because of this, they would not have reliable measures of these three aspects of fluid reasoning ability (*Gf*), but only a composite reasoning factor. They would thus see only evidence for *g* or *Gf* and not for the distinguishably different abilities to reason with words (as well as the concepts they can signify), with numbers or symbols (as well as the concepts they can signify), and stylized spatial figures (as well as the concepts they can signify). The assertion that nonverbal figural reasoning tests are fair proxies for verbal or quantitative reasoning tests rests on the mistaken assumption that, absent task-specific factors, all reasoning tests measure more or less the same thing.

Table 1 shows why this assumption is untenable. The table shows the percentage of high-scoring students in the 2000 CogAT standardization sample who had different score profiles on the CogAT multilevel battery. The most recent edition of CogAT reports a profile score for each student, which summarizes the level and pattern of his or her scores across the verbal, quantitative, and nonverbal reasoning batteries. Example profiles are 3A, 9B(V-), and 6C(V+Q-). The number is the student's median age stanine on the three batteries. Stanines range from 1 (lowest 4% of scores in the distribution) to 9 (highest 4% of scores in the distribution). The median stanine estimates the overall level of the profile. The first letter tells whether all three scores were at the same level (an "A" profile), whether one score was above or below the other two scores (a "B" profile), or whether two scores showed a significant contrast (a "C") profile. In the examples above, 3A means that the median age stanine was 3 and that the three scores did not differ significantly from one another.[9] The second example, 9B(V-), means that the median age stanine was 9 and that the score on the Verbal Battery was significantly lower than the scores on the Quantitative and Nonverbal batteries. The last profile, 6C(V+Q-), shows a relative strength on the Verbal Battery and relative weakness on the Quantitative Battery. Finally, in an effort to call attention to unusually large differences, profiles with scores that differ more than 24 points on the SAS scale[10] are all labeled E (for "extreme"). For example, 8E(N-) means that the median stanine was 8 and that the score on the Nonverbal Battery was at least 24 points lower than the score on one of the other two batteries.

Given the interest here in identifying gifted students, only scores for the 11,031 students who had a median stanine of 8 or 9 were included in the data summarized in the table.[11] This represents the top 10–11% of students in the national sample. If all three of these highly reliable[12] reasoning scores measure approximately the same thing, then the majority of students—especially White students—should have approximately equal scores on the Verbal, Quantitative, and Nonverbal batteries. Here, this would be represented by an "A" profile. On the contrary, only 42% of high-scoring White students showed this profile. Stated the other way, the majority of high-scoring White students showed significantly uneven profiles of reasoning abilities. Of this majority, 28.9% showed a significant, but not extreme, strength or weakness in one area (see the "Total B" row). Another 13.6% showed an extreme strength or weakness (see the "Total E_B" row). A relative weakness was much more common than a relative strength.[13] Finally, 15.4% showed a significant (12.3%) or extreme (3.1%) contrast between two scores ("Total C" and "Total E_C" rows). Clearly, one size does not fit all. Giftedness in reasoning abilities is multidimensional, not unidimensional (see Achter, Lubinski, & Benbow, 1996, for a similar conclusion).

Table 1

Percent of High–Scoring Students (Median Stanine = 8 or 9) Showing Different Profiles of Verbal,
Quantitative, and Nonverbal Reasoning Abilities on the CogAT Form 6 Multilevel Battery

Profile	White	Black	Hispanic	Asian American	Indian	Other or Missing	Total
				Ethnicity			
All scores at the same level							
A	42.0	28.5	31.8	30.1	38.5	37.1	40.4
One score above or below							
B (V+)	2.6	1.7	2.1	1.2	1.0	1.3	2.4
B (V−)	9.1	11.9	14.1	15.1	11.4	7.7	9.7
B (Q+)	2.6	1.1	2.8	2.9	1.4	4.4	2.6
B (Q−)	6.1	8.6	5.1	4.4	8.2	4.9	6.1
B (N+)	2.2	2.1	2.0	2.2	0.0	3.2	2.2
B (N−)	6.3	13.5	4.9	6.7	8.3	7.1	6.5
Total B	28.9	39.0	30.9	32.6	30.3	28.7	29.5
Extreme B profile							
E (V+)	1.4	0.0	0.7	1.8	0.0	0.1	1.3
E (V−)	4.3	8.6	11.4	13.8	4.3	7.6	5.3
E (Q+)	1.5	1.4	1.5	1.3	2.0	1.8	1.5
E (Q−)	2.5	4.1	0.4	0.3	3.7	2.8	2.4
E (N+)	1.4	0.0	1.8	2.1	1.7	2.9	1.5
E (N−)	2.5	5.8	2.5	1.7	1.4	2.0	2.5
Total E$_B$	13.6	20.0	18.3	21.0	13.1	17.3	14.4
Two scores contrast							
C (V+Q−)	2.3	0.8	2.5	1.1	0.7	2.3	2.2
C (V−Q+)	2.0	1.6	2.6	2.0	5.5	2.5	2.1
C (V+N−)	2.2	0.8	2.2	0.3	4.3	1.2	2.1
C (V−N+)	2.1	1.0	3.3	2.8	0.7	1.0	2.1
C (Q+N−)	1.7	2.4	1.4	1.5	1.2	2.0	1.7
C (Q−N+)	2.1	0.8	0.7	1.2	2.7	4.1	2.0
Total C	12.3	7.5	12.8	8.9	15.0	13.2	12.1
Extreme C profile							
E (V+Q−)	0.7	0.0	0.3	0.3	0.5	0.0	0.6
E (V−Q+)	0.5	1.4	3.6	3.7	0.9	1.1	0.8
E (V+N−)	0.4	3.5	0.0	1.0	0.0	0.4	0.5
E (V−N+)	0.7	0.1	2.1	1.7	0.7	1.0	0.8
E (Q+N−)	0.5	0.0	0.0	0.7	0.0	1.2	0.5
E (Q−N+)	0.2	0.0	0.1	0.0	1.0	0.0	0.2
Total E$_C$	3.1	5.1	6.2	7.4	3.1	3.7	3.5
N[a]	9,361	176	317	550	195	70	11,031

Note. All columns total 100. V = Verbal; Q = Quantitative; N = Nonverbal; A = All three scores at approximately the same level; B = One score above or below the other two scores; C = Two scores contrast significantly; E = Scores differ by at least 24 SAS points.
[a]*Frequencies based on the weighted data. N count shows the actual number of cases.*

The profiles for minority students are even more interesting. If tests with verbal and quantitative content are inherently biased against minority students, then there should be very few students with an even or "A" profile. Most should show an N+ profile (i.e., a much higher score on the nonverbal battery than on the verbal and quantitative batteries). On the contrary, approximately 30% of the Black, Hispanic, and Asian students also showed an even profile across the three batteries. Neither N+ nor E(N+) profiles occurred with greater frequency for these students than for White students. As expected, V- profiles were more common for minority students. Note, however, that 13.4% of the White students also showed either a significant (9.1%) or extreme (4.3%) V- profile. Further, Black students were much more likely than other ethnic groups to show a significantly lower score on the nonverbal battery (an N- profile) than on either the Verbal or Quantitative batteries. Fully 19.3% showed a significant (13.5%) or extreme (5.8%) relative weakness on the Nonverbal Battery. *This means that screening students with a nonverbal reasoning test will actually eliminate many of the most academically capable Black students in the sample.* Indeed, the only extreme profile that was more common for Black students was a verbal strength coupled with a nonverbal weakness, E(V+N-). For Hispanic and Asian American students, the most common extreme contrast profile was a verbal weakness coupled with a quantitative strength, E(V-Q+). Once again, this argues for the importance of estimating the quantitative reasoning abilities of minority students.

Spatial Strengths as Inaptitude for Academic Learning?

Although figural reasoning ability is not the same as spatial ability, the two constructs fall in the same branch of a hierarchical model of abilities (Gustafsson & Undheim, 1996) or in the same slice of the radex model (Marshalek, Lohman, & Snow, 1983). In both of these models, figural reasoning abilities are closer to *g*. Spatial abilities, although still highly *g*-loaded, fall further down in a multilevel hierarchical model or somewhat further from the center of the radex. The key difference is that figural reasoning tests require examinees to make inferences, deductions, and extrapolations from figural stimuli, whereas spatial tests require the ability to create images that preserve configural information in the stimulus—often while performing analog transformations of those images. Many figural tests, of course, sample both reasoning and spatial processing, depending on how items are constructed and how examinees choose to solve them.

These distinctions become important in trying to understand one of the most unexpected findings in our analyses of the CogAT standardization data. At all levels of achievement in grades 3–12, students who showed a relative strength on the CogAT Nonverbal Battery showed lower achievement in some areas than students who had the same levels of verbal and quantitative abilities, but a relative weakness on the Nonverbal Battery. In other words, a relative strength in nonverbal reasoning seems to be an *inaptitude* for some aspects of school learning—particularly the sorts of basic skills students must learn in elementary school. The effect was particularly strong on verbal achievement in domains such as spelling and language usage at the elementary school level and for students who scored in the average range (Lohman & Hagen, 2001c, p. 102). But, the effect was clearly present among the most able students, as well (p. 105), and in other achievement domains (e.g., Vocabulary, Reading Comprehension, Math Computation, and Composite Achievement Score). The only subtest of the Iowa Tests of Basic Skills (ITBS) on which students with an N+ profile consistently outperformed those with an N- profile was on the Maps and Diagrams test.

There are several reasons why this could be the case. One possibility is that students with an N+ profile perform especially well on figural reasoning tests because they have unusually strong spatial abilities. Such students may well find themselves mismatched in an educational system that requires mostly linear and verbal modes of thinking, rather than their preferred spatial modes of thinking (Lohman, 1994). Another possibility is that achievement tests generally do not measure spatial modes of thinking. Grades or other measures of accomplishment in literature, science, or mathematics may not show such effects. However, Gohm, Humphreys, and Yao (1998) found that students gifted in spatial ability underperformed on a wide range of school interest and achievement measures that included both tests and grades. Although one could envision an alternative educational system in which this might not be the case, underperformance cannot be attributed to the verbal bias of the achievement tests. A third possibility is that a high nonverbal score reflects a strength in fluid reasoning ability, rather than in spatial ability. Students who excel in fluid (as opposed to crystallized) abilities are particularly adept at solving unfamiliar problems, rather than the more familiar sort of problems routinely encountered in school. However, if this were the case, deficits in mathe-

matics should be as common as deficits in the more verbal domains. High spatial abilities, on the other hand, are commonly linked to problems in verbal fluency, spelling, and grammar (Shepard, 1978). Thus, the effect seems more plausibly linked to a preference for spatial thinking, rather than to a relative strength in fluid reasoning. Finally, one might hypothesize that the effect in undifferentiated samples reflects the performance of minority students. Such students would be particularly likely to underachieve on verbal achievement tests that emphasize specific language skills such as spelling, grammar, and usage. However, our analyses show that these effects are even stronger for minority students than for White students.[14] *This means that selecting students on the basis of their nonverbal reasoning abilities without also attending to their verbal and quantitative reasoning abilities will select some students who are even less likely to achieve at high levels than students with much lower nonverbal reasoning scores*. Notice that an isolated strength in nonverbal reasoning is not the same thing as strengths in both quantitative and nonverbal reasoning or in both verbal and nonverbal reasoning or in all three. Students with these score profiles do not show the deficits observed in the N+ group. This concurs with the finding of Humphreys, Lubinski, and Yao (1993) that engineers were more likely to excel on both spatial and mathematical abilities. However, unless these other reasoning abilities are measured, one has no way of knowing whether a particular student with a high nonverbal score is even less likely than other students to achieve well.

Reconceptualizing Potential as Aptitude

The primary purpose of schooling is to assist students in developing expertise in particular domains of knowledge and skill that are valued by a culture. The primary purpose of programs for the gifted and talented ought to be to provide appropriate levels of challenging instruction for those students who have demonstrated high levels of accomplishment in one or more of these domains. This can be done through acceleration or advanced placement, for example. The secondary purpose of such programs ought to be to provide enrichment or intensive instruction for those who show potential for high levels of accomplishment. These students commonly need different levels of challenge than those who have already demonstrated high levels of competence in a domain. Measuring accomplishment is difficult. Measuring potential for accomplishment is even more difficult; more troubling, it is fraught

with misconceptions and pitfalls. For example, some misconstrue potential as latent or suppressed competence waiting to burst forth when conditions that prevent its expression are removed (Humphreys, 1973). Such misconceptions have prompted others to reject potential as a pie-in-the-sky concept that refers to the level of expertise an individual might develop if he or she were reared in some mythically perfect environment. A more moderate position is to understand potential as readiness to acquire proficiency in some context—that is, as aptitude.

A Definition of Aptitude

Students approach new educational tasks with a repertoire of knowledge, skills, attitudes, values, motivations, and other propensities developed through life experiences to date. The school experience may be conceptualized as a 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. Specifically, 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 individuals bring to a situation, the few that assist them in performing well in that situation function as aptitudes. Examples 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. Aptitudes for learning thus go beyond cognitive abilities. Aspects of personality and motivation commonly function as aptitudes, as well.

Prior achievement is commonly an important aptitude for future achievement. Whether prior achievement functions as aptitude in a particular situation depends on both the person's propensity to use prior knowledge in new situations and the demand and opportunity structure of the situation. Therefore, understanding which of an individual's characteristics are likely to function as aptitudes begins with a careful examination of the demands and affordances of the target environment. In fact, defining the situation is part of defining the aptitude (Snow & Lohman, 1984). The affordances of an environment are what it offers or 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 affords the use of reasoning abilities; direct instruction often does not.

The second step is to identify those characteristics (or propensities) of individuals that are coupled with task or situation affordances. 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. Although there is much data that confirms this assertion, a simple thought experiment will illustrate it. Consider the likelihood that a second grader with high scores on a nonverbal reasoning test, but no background in mathematics will succeed in learning arithmetic. Now, consider the likelihood that a graduate student with equally high scores on a nonverbal reasoning test, but no background in mathematics will succeed in a graduate-level mathematics class. Knowledge matters.

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 evaluation of essays, speeches, drawings, or science experiments produced by others (Rodriguez, 2003).

The second most important learner characteristic for academic learning is 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 they signify and (b) with quantitative symbols and the concepts they signify. Thus, the critical reasoning abilities for all students (minority and majority) are verbal and quantitative. Figural reasoning abilities are less important and thus show lower correlations with school achievement.

The Relative Importance of Prior Achievement and Reasoning Abilities in the Prediction of Future Achievement

Evidence for these claims about the relative importance of prior knowledge and skill versus the ability to reason in different symbol systems is shown in Tables 2 and 3. Students in a large midwestern school district were retested with different levels of CogAT and the ITBS in grades 4, 6, and 9. The data in Tables 2 and 3 are for the 2,789 students who had scores for both the grade 4 and grade 9 test administrations and the 4,811 students who had scores for both the grade 6 and the grade 9 testings. The dependent variable in Table 2 is grade 9 Reading Scale Score; the dependent variable in Table 3 is grade 9 Mathematics Scale Score.

The critical question here is whether prior achievement and prior ability both contribute to the prediction of grade 9 achievement. This was addressed in a series of multiple regression analyses. The independent variables were entered in two blocks: Block 1 contained the prior achievement test scores for reading, language, and mathematics; CogAT scores for verbal, quantitative, and nonverbal reasoning; and sex. Block 2 contained the interactions between each of the six test scores and sex. These interaction terms test whether the prediction equations differed significantly for males and females.

Look first at the prediction of grade 9 reading achievement in Table 2. The first column shows correlations between the grade 4 achievement and ability test scores and the grade 9 reading scale score. CogAT Verbal Reasoning had the highest correlation ($r = .741$) followed by grade 4 reading achievement ($r = .732$). The regression analysis, however, shows that grade 4 reading achievement was the relatively stronger predictor of grade 9 reading achievement when all scores were considered simultaneously ($\beta = .359$ vs. $\beta = .288$).

The right side of Table 2 shows that grade 6 reading achievement was also the best predictor of grade 9 reading achievement. However, both reading achievement and CogAT Verbal interacted with sex in the grade 6 predictor. The within-sex regressions showed that, for boys, grade 6 reading achievement was a slightly better predictor than CogAT Verbal, whereas for girls the situation was reversed.[15]

Table 3 shows a similar pattern of results for the prediction of grade 9 mathematics from grade 4 achievement and ability scores (left panel) and grade 6 achievement and ability scores (right panel). At grade 4, mathematics achievement was the strongest contributor to the prediction ($\beta = .23$), whereas at grade 6, CogAT Quantitative reasoning predominated ($\beta = .262$). Once again, both of these variables interacted with sex. This time grade 6 mathematics achievement had the largest beta weight for girls, whereas for boys, CogAT Quantitative was largest.

Correlations among the independent variables make these sorts of comparisons among beta weights suggestive, rather than definitive (Pedhazur, 1982). Nonetheless, it is clear that the two most important predictors of future reading (or math) achievement are cur-

Table 2

Prediction of Grade 9 Reading Achievement from Grade 4 (N = 2,789)
or Grade 6 (N = 4,811) Achievement and Ability Scores

	Grade 4				Grade 6			
	r	b	β	p	r	b	β	p
Constant		36.454		0.000*		28.948		0.000*
ITBS Reading (R)	0.732	0.556	0.359	0.000*	0.797	0.558	0.437	0.000*
ITBS Language (L)	0.642	0.010	0.007	0.148	0.684	0.059	0.055	0.005*
ITBS Mathematics (M)	0.635	0.196	0.126	0.000*	0.657	0.046	0.037	0.078
CogAT Verbal (V)	0.741	0.617	0.288	0.000*	0.796	0.722	0.321	0.000*
CogAT Quantitative (Q)	0.595	0.066	0.027	0.912	0.643	-0.006	-0.003	0.901
CogAT Nonverbal (N)	0.576	0.128	0.053	0.000*	0.602	0.114	0.048	0.005
Sex (S)[a]	0.075	-5.157	-0.073	0.038	0.054	1.904	0.027	0.719
Interactions With Sex								
R x S		-0.025	-0.071	0.694		-0.075	-0.246	0.047*
L x S		0.065	0.184	0.294		-0.020	-0.066	0.507
M x S		-0.062	-0.173	0.330		0.019	0.060	0.613
V x S		0.053	0.078	0.615		0.165	0.244	0.031*
Q x S		-0.131	-0.191	0.206		0.082	0.123	0.240
N x S		0.187	0.282	0.031*		-0.072	-0.111	0.233

Note. R^2 = .621 for Grade 4; and R^2 = .702 for Grade 6. r = Pearson product-moment correlation; b = unstandardized regression coefficient; β = standardized regression coefficient; p = probability. Achievement scores are from levels 10 (grade 4), 12 (grade 6), and 14 (grade 9) of the ITBS Survey Battery, Form K (Hoover, Hieronymous, Frisbie, & Dunbar, 1993). Ability scores are from levels B (grade 4) and D (grade 6) of CogAT, Form 5 (Thorndike & Hagen, 1993).
[a] *Girl = 1*
★ p < .05

rent reading (or math) achievement and verbal (or quantitative) reasoning ability. The fact that prior achievement and reasoning ability are the two largest contributors to the prediction of grade 9 achievement runs counter to assertions that verbal (or quantitative) reasoning tests measure the same thing as verbal (or math) achievement tests. Rather, these results support the claim that the two most important aptitudes for academic learning are current achievement in the domain and domain-specific reasoning ability.

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 considered. Our data suggest that the two should be weighted approximately equally, although the relative importance of prior achievement and abstract reasoning will depend on the demands and affordances of the instructional environment and 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 testings or when content changes abruptly (as from arithmetic to algebra), then reasoning abilities become more important. Novices typically rely more on knowledge-lean reasoning abilities than do domain experts. Because children are universal novices, reasoning abilities are therefore more important in the identification of academic giftedness in children, whereas evidence of domain-specific accomplishments is relatively more important for adolescents.

The Prediction of Achievement for Minority Students

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

Table 3

Prediction of Grade 9 Mathematics from Grade 4 (N = 2,789)
or Grade 6 (N = 4,811) Achievement and Ability Scores

	Grade 4				Grade 6			
	r	*b*	β	*p*	*r*	*b*	β	*p*
Constant		42.692		0.000*		39.259		0.000*
ITBS Reading (R)	0.585	0.206	0.126	0.000*	0.664	0.207	0.155	0.000*
ITBS Language (L)	0.574	0.024	0.014	0.627	0.662	0.019	0.017	0.428
ITBS Mathematics (M)	0.683	0.377	0.231	0.000*	0.743	0.270	0.209	0.000*
CogAT Verbal (V)	0.665	0.255	0.113	0.002*	0.712	0.308	0.131	0.000*
CogAT Quantitative (Q)	0.672	0.504	0.194	0.000*	0.746	0.619	0.262	0.000*
CogAT Nonverbal (N)	0.637	0.452	0.178	0.000*	0.670	0.309	0.124	0.000*
Sex (S)[a]	-0.099	-1.234	-0.098	0.467	-0.096	-6.858	-0.092	0.252
Interactions With Sex								
R x S		-0.111	-0.302	0.116		-0.109	-0.343	0.011*
L x S		-0.003	0.069	0.963		-0.004	-0.013	0.906
M x S		0.018	0.048	0.800		0.092	0.280	0.028*
V x S		0.172	0.241	0.141		0.100	0.141	0.246
Q x S		-0.073	-0.101	0.525		-0.035	-0.050	0.658
N x S		0.092	0.132	0.338		0.003	0.004	0.969

Note. R^2 = .578 for Grade 4; and R^2 = .654 for Grade 6. r = Pearson product-moment correlation; b = unstandardized regression coefficient; β = standardized regression coefficient; p = probability. Achievement scores are from levels 10 (grade 4), 12 (grade 6), and 14 (grade 9) of the ITBS Survey Battery, Form K (Hoover, Hieronymous, Frisbie, & Dunbar, 1993). Ability scores are from levels B (grade 4) and D (grade 6) of CogAT, Form 5 (Thorndike & Hagen, 1993).
[a] Girl = 1
＊p < .05

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.

Table 4 shows an example for predicting Reading Total on the ITBS in grades 1–8 and on the Iowa Tests of Educational Development (ITED) in grades 9–12. The predictors were CogAT Verbal, Quantitative, and Nonverbal SAS scores. Grades 1 and 2 used the CogAT Primary Battery, whereas grades 3–12 used one of the eight levels of the CogAT Multilevel Battery.

The left half of the table reports the analyses for all students in the sample; the right half of the table shows the same analyses for the Hispanic students. Each row first

reports the raw correlations between the three CogAT SAS scores and Reading Total. Next, the results of a multiple regression in which Reading Total was predicted from the three CogAT scores are reported. Entries in this portion of the table are the standardized regression coefficients (beta weights) and the multiple correlations.

Look first at the last row in the table, which reports the average entry in each column. The average correlations between the three CogAT scores and reading achievement were .78, .65, and .60 for Verbal, Quantitative, and Nonverbal reasoning. The multiple correlation between all three tests and reading achievement was .80, which is substantially higher than the correlation of .60 observed for the Nonverbal Battery. The good news, then, is that the score on the Nonverbal Battery predicts reading achievement. The bad news is that the prediction is relatively poor when compared to the multiple correlation across all three batteries.

For Hispanics, the correlations with reading achievement were .77, .62, and .53 for verbal, quantita-

Table 4

Prediction of ITBS Form A (grades 1–8) or ITED Form A (grades 9–12) Reading Total Scale Score From CogAT Form 6 Verbal (V), Quantitative (Q), and Nonverbal (NV) Reasoning Standard Age Scores for All Students and for Hispanic Students by Grade

	All Students								Hispanic Students							
	Correlation				β				Correlation				β			
Grade	V	Q	NV	N	V	Q	NV	Mult. R	V	Q	NV	N	V	Q	NV	Mult. R
1	.56	.60	.52	11,424	.216	.311	.193	.64	.59	.61	.46	1,051	.271	.314	.139	.64
2	.69	.66	.55	11,882	.432	.229	.154	.73	.65	.62	.47	1,093	.424	.237	.114	.69
3	.80	.66	.62	13,678	.669	.138	.052	.81	.80	.63	.58	1,382	.714	.121	-.009	.80
4	.80	.66	.62	13,630	.660	.137	.066	.81	.80	.60	.56	1,462	.726	.070	.043	.81
5	.81	.66	.62	13,935	.696	.107	.057	.82	.79	.62	.55	1,423	.720	.061	.046	.80
6	.83	.67	.62	13,811	.710	.120	.049	.84	.83	.64	.55	1,199	.759	.122	-.019	.84
7	.84	.69	.63	13,164	.714	.150	.023	.85	.81	.65	.55	1,129	.709	.155	-.008	.82
8	.84	.67	.63	11,178	.730	.115	.041	.85	.81	.62	.57	1,095	.699	.089	.079	.82
9	.82	.67	.63	8,112	.693	.126	.053	.83	.79	.60	.50	779	.727	.100	-.005	.80
10	.82	.66	.60	6,083	.706	.128	.042	.83	.83	.84	.64	453	.760	.115	.004	.85
11	.79	.62	.57	5,078	.686	.133	.021	.80	.80	.59	.53	382	.688	.081	.107	.81
12	.78	.60	.59	4,106	.670	.098	.072	.79	.78	.66	.56	252	.602	.234	.036	.80
Mean	.78	.65	.60	10,507	.632	.149	.069	.80	.77	.62	.53	975	.650	.142	.044	.79

Note. Correlations are with Reading Total. β = standardized regression coefficient. Mult. R = multiple correlation. Reading achievement scores are from the ITBS, Form A (Hoover, Dunbar, & Frisbie, 2001) at grades 1–8 and from the ITED, Form A (Forsyth, Ansley, Feldt, & Ahot, 2001) at grades 9–12. CogAT scores are from Form 6 (Lohman & Hagen, 2001a).

T a b l e   5

Increment in R–Square Observed when CogAT Verbal, Quantitative, or Nonverbal Scores are Added Last to the Prediction of Reading Achievement (Left Panel) or Mathematics Achievement (Right Panel)

Grade	Reading Total			Mathematics Total		
	Verbal	Quantitative	Nonverbal	Verbal	Quantitative	Nonverbal
CogAT Form 6 Primary Battery						
1	0.019	0.032	0.021	0.018	0.085	0.018
2	0.079	0.018	0.013	0.007	0.093	0.022
CogAT Form 6 Multilevel Battery						
3	0.191	0.007	0.001	0.033	0.071	0.009
4	0.191	0.007	0.002	0.023	0.077	0.011
5	0.206	0.004	0.001	0.022	0.080	0.011
6	0.225	0.005	0.001	0.023	0.091	0.009
7	0.219	0.008	0.000	0.017	0.094	0.008
8	0.234	0.004	0.001	0.024	0.090	0.007
9	0.208	0.005	0.001	0.031	0.085	0.005
10	0.234	0.006	0.001	0.031	0.095	0.008
11	0.233	0.007	0.000	0.035	0.093	0.006
12	0.221	0.004	0.002	0.028	0.097	0.008

Note. Reading (or Mathematics) Total is from Form A of the ITBS (Hoover et al., 2001) at grades 1-8 and Form A of the ITED (Forsyth et al., 2001) at grades 9–12. CogAT scores are from Form 6 (Lohman & Hagen, 2001a). Sample sizes at each grade are reported in column 5 of Table 4.

tive, and nonverbal reasoning. If any thing, then, the nonverbal score was less predictive of reading achievement for Hispanics than for Whites. The standardized regression coefficients mirrored this pattern: Verbal reasoning ability was the best predictor of reading achievement for Hispanic students (β = .650); nonverbal reasoning was the worst (β = .044). Indeed, sometimes the nonverbal score had a negative weight.

The unique contributions of verbal, quantitative, and nonverbal reasoning abilities to the prediction of achievement were also examined in separate set of regression analyses. The question was this: What does each of these three reasoning abilities add to the prediction of reading or math achievement once the other two abilities are taken into account? The critical statistic is the increment in R-square that is observed when the third predictor is added to the regression. The results for Reading Total are shown in the left half of Table 5 and for Mathematics Total in the right half of Table 5.

For reading achievement, the verbal reasoning score added enormously to the prediction even after the quan-

titative and nonverbal scores were entered into the equation. The median increment in R^2 for the Multilevel Battery (grades 3–12) was 0.22. When entered last, quantitative reasoning added a small, barely significant increment of .006 to the R-square. Finally, when entered last, nonverbal reasoning made a significant contribution only for the orally administered Primary Battery (grades 1 and 2). For grades 3–12, nonverbal reasoning contributed only .001 to the R^2.

A similar, but less lopsided set of results were obtained for the prediction of mathematics achievement. As expected, quantitative reasoning made the largest unique contribution. The median increment in R^2 was .091 at grades 3–12. The corresponding increments in R^2 for verbal reasoning and nonverbal reasoning were .026 and .008, respectively.

Therefore, whether judged by the relative magnitudes of the beta weights (Table 4) or the increment in R-square when entered last (Table 5), the nonverbal score was clearly much less important than the verbal and quantitative scores for the prediction of achievement.

Further, at least for reading achievement, the regression equations were essentially the same for Hispanics as for the total sample. But, do these results generalize to other ethnic groups and other domains of academic achievement?

This question was addressed in a final set of regression analyses that included only Whites and Hispanics, or Whites and Blacks, or Whites and Asian Americans. As in Table 3, a variable for ethnicity was also coded (e.g., 0 = White, 1 = Hispanic) and then interactions between each of the three CogAT scores and ethnicity were computed. The interaction terms test the hypothesis that the regression weights for one or more of the CogAT scores differ for the ethnic groups being compared. The median increment in R-square that was observed when all the three interaction terms were added to the model was 0.0% of the variance for the White-Hispanic analyses, 0.1% for the White-Black analyses, and 0.0% for the White-Asian analyses. In other words, the regression equations that best predict reading, mathematics, social studies, and science achievement in grades 1–12 for Hispanic, Black, and Asian American students are the same as the regression equations that best predict the performance of White students in each of these domains at each grade.

Verbal Reasoning Abilities of ELL Students

Although the predictors of achievement are the same for White, Hispanic, Black, and Asian American students, some would argue that tests that make any demands on students' language abilities are unfair to those who do not speak English as a first language or who speak a dialect of English not spoken by most test takers. Put differently, how can we estimate the verbal (and, to a lesser extent, quantitative) reasoning abilities of students who speak English as a second language? Although quantitative reasoning tests often make few language demands, verbal reasoning tests typically measure the ability to reason using the dialect of the English language commonly used in formal schooling. However, even though this is not the same as the ability to reason in another language, verbal abilities in any language seem to depend on a common set of cognitive processes. This is shown most clearly in studies of bilingual students in which the predictors of achievement are largely the same not only across ethnic groups, but also within a bilingual population across different languages (Gustafsson & Balke, 1993; Lindsey, Manis, & Baily, 2003). Put differently, the problem of identifying those Hispanic students best able to reason in English is similar to the problem of identifying those English-speaking students most likely to succeed in understanding, reading, and writing in French. (I leave out speaking skills because language production abilities seem to involve specific abilities unrelated to academic competence [Carroll, 1981].) The evidence is clear that competence in understanding, reading, and writing English are much better predictors of success in learning to understand, read, and write French than are numerical reasoning or nonverbal reasoning (Carroll). But, the best prediction is given by the ability to comprehend and reason in French after some exposure to the French language.

The same logic applies to the problem of identifying bilingual students who are most likely to achieve at high levels in domains that require verbal reasoning abilities, but in English. A test that assessed the abilities of Hispanic students to reason in Spanish would predict their ability to reason in English. But, the best and most direct prediction is given by a test that measures how well they have learned to reason in English, given some years of exposure to the language.

Notice that an aptitude perspective requires that one be much more specific about the demands and affordances of the learning situation than a model that presumes that an undifferentiated g should best predict performance in any context. In particular, success in 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. Failure to measure these abilities does not somehow render them any less important. There is no escaping the fact that the bilingual student most likely to succeed in school will exhibit strong verbal reasoning skills in his or her first language and, even more importantly, in the language of instruction. Thus, an aptitude perspective leads one to look for those students who have best developed the specific cognitive (and affective) aptitudes most required for developing expertise in particular domains. For example, the Black, Hispanic, or Asian American students most likely to develop expertise in mathematics are those who obtain the highest scores on tests of mathematics achievement and quantitative reasoning. Identifying such students requires this attention to proximal, relevant aptitudes, not distal ones that have weaker psychological and statistical justification.

One Norm for All?

But, how can schools take into account the relevant aptitudes and at the same time increase the diversity of the selected group? On average, Hispanic and Black students score below White students on both measures of achievement and reasoning ability. For Blacks, the problem is not only lower mean score, but also a smaller variance. In their review of the Black-White test score differences on six large national surveys, Hedges and Nowell (1998) found that the variances of the score distributions for Black students were 20–30% smaller than the variances of the score distributions for Whites. The combination of a lower mean and a smaller variance means that very few Blacks obtain high scores in the common score distributions (see Koretz, Lynch, & Lynch, 2000).

Because of this, schools have looked for measures of accomplishment or ability that show smaller group differences. One can reduce the differences by careful construction of tests. Most well-designed ability and achievement tests go to great lengths to remove irrelevant sources of difficulty that are confounded with ethnicity. But, one cannot go very far down this path without getting off the path altogether. For example, one can use teacher ratings of student creativity instead of measures of achievement or, in the case of ability tests for ELL students, measures of figural reasoning ability instead of verbal and quantitative reasoning abilities. Creativity is a good thing, but it is not the same thing as achievement. Schools should aim to develop both. However, one cannot substitute high creativity for high achievement. Further, ratings of creativity are much less reliable than measures of achievement. Group differences will always be smaller on a less reliable test than a more reliable test. In the extreme, a purely random selection process would show no group differences. Similarly, a less valid predictor of achievement (such as a figural reasoning test) may be successful in identifying relatively more minority students, but more of these will be the wrong students (see Figure 1). This should concern everyone, especially the minority communities who hope that students who receive the special assistance offered in programs for the gifted and talented will someday become the next generation of minority scholars and professionals. Getting the right kids is much more important than getting the right number of kids.

The problem, I believe, is that programs for the gifted and talented have not clearly distinguished between the criteria that should be used to identify students who currently display extraordinary levels of academic accomplishment from the criteria that should be used to identify those whose current accomplishments are lesser, but who show potential for developing academic excellence (see Lohman, in press, for an elaboration of this argument). In identifying students whose current accomplishments are extraordinary, common measures and common criteria are most appropriate. A third-grade student who will be studying algebra needs a level of mathematical competence that will allow him or her to succeed in the algebra class. Other aptitudes are important, but none could compensate for the lack of requisite mathematical knowledge.

Potential for developing a high level of accomplishment, on the other hand, is a much slipperier concept. Even in the best of cases, predictions of future achievement are often wrong. For example, Table 2 showed the prediction of grade 9 reading achievement from six measures of achievement and reasoning abilities obtained in grade 4, plus sex, and the six interactions between sex and each of the achievement and ability tests. The R-square in this analysis was .621, and so the R was .788. Although this is a substantial correlation, it means that only slightly more than half of the students who were predicted to be in the top 10% of the grade 9 reading achievement on the basis of their gender, grade 4 achievement, and grade 4 ability test scores actually obtained grade 9 reading scores at this level.[16] Put differently, even with excellent measures of prior achievement and ability, we could forecast whether a child would fall in the top 10% of the distribution 5 years later with only slightly more than 50% accuracy. Furthermore, these predictions are even less likely to hold if schools adopt interventions that are designed to falsify them. For example, intensive instruction in writing for students who have had little opportunity to develop such skills on their own can markedly change the relationship between pretest and posttest scores. Differences in opportunity to learn can be substantial and must be taken into account when making inferences about potential to acquire competence. This is recognized in the U.S. Department of Education's guidelines for identifying gifted students. There, gifted children are defined as those who "perform or show the potential for performing at . . . high levels of accomplishment *when compared with others of their age, experience, or environment*" (U.S. Department of Education, 1993; emphasis added). The definition confounds performance with potential for performance. I would argue that, although current levels of accomplishment should be judged using standards that are the same

for all, potential for acquiring competence must always be made relative to circumstances. To do otherwise presumes that abilities can be assessed independently of opportunities to develop them. This is not possible.

Therefore, when estimating a student's potential to acquire competence, schools cannot blindly apply a uniform cutoff score, however fair such a procedure may seem or administratively convenient it may be. The 10-year-old ELL student with, say, 3 years of exposure to the English language who has learned to reason with English words at the same level as the average 10-year-old native speaker has exhibited much greater potential for language learning than the 10-year-old native speaker. Schools can best identify such bilingual students by examining frequency distributions of scores on the relevant achievement and reasoning tests. It is helpful to know where students stand in relation to all of their age or grade peers, as well to as those with similar backgrounds. *Concretely, the Hispanic student with the highest probability of developing academic excellence in a particular domain is the student with the highest achievement in that domain and the best ability to reason in the symbol systems most demanded for new learning in that domain.*

In general, judgments about potential to acquire proficiency have both empirical and ethical dimensions, both of which need to be addressed. The empirical question is "Who is most likely to attain academic excellence if given special assistance?" The ethical question is "Who is most likely to need the extra assistance that schools can provide?" For some students, special programs to develop talent provide an ancillary opportunity; for other students, they provide the only opportunity.

Once these high-potential students have been identified, the next step is to intervene in ways that are likely to assist them in developing their academic skills. This involves much more than a short pull-out program. Students, their parents, and their teachers need to understand that the development of high levels of academic competence involves the same level of commitment (and assistance) as does the development of high levels of athletic or musical competence. Being identified as having the potential to achieve at a high level should not be confused with achieving at a high level. One possibility is to solicit the active involvement of high-achieving minority adults to work with students in developing their academic expertise. I am particularly impressed with programs such as Urban Debate League in Baltimore (see http://www.towson.edu/news/campus/msg02628.html). Some of the many good features of such programs are an emphasis on production (rather than reception) of language, teamwork, long hours of guided practice, apprenticeship programs, competitions that motivate, and the active involvement of adults who serve as role models.

The Proper Role of Nonverbal Ability Tests

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 everyone's height just because we find it difficult to measure some people's weight. 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. The trouble, however, is that minority or majority students with the highest nonverbal reasoning test scores are not necessarily the students who are most likely to show high levels of achievement—either currently or at some later date. Rather, those students with the highest domain-specific achievement and who reason best in the symbol systems used to communicate new knowledge are the ones most likely to achieve at a higher level. Therefore, high scores on the nonverbal test should *always* be accompanied by evidence of high (but not necessarily stellar) accomplishment in a particular academic domain or by evidence that the student's verbal or quantitative reasoning abilities are high relative to those in similar circumstances. Most schools have this evidence for achievement, and those that administer ability tests that contain scores for verbal and quantitative reasoning in addition to the nonverbal score would have the corresponding evidence for ability, as well. For many ELL students, mathematics achievement, quantitative reasoning abilities, or both 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. This was shown by the high frequency of N– profiles for Black students in Table 1. Thus, less-than-stellar performance on the nonverbal test is even less informative for these students than for other students.

Absent ancillary information on verbal or quantitative abilities and achievement, then, 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 the nonverbal test are thus a useful

supplement. They sometimes add to the prediction of achievement—especially in the quantitative domains. Thus, 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. 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 nonverbal test may appear to reduce bias, but when used alone it actually increases bias by failing to select those most likely to profit from enrichment.

Summary

Measures of academic accomplishment (which include, but are not limited to, norm-referenced achievement tests) should be the primary criteria for defining academic giftedness. Such assessments not only measure knowledge and skills that are important aptitudes for academic learning, but they also help define the type of expertise schools aim to develop. Because of this, they can direct the efforts of those who would aspire to academic excellence. When properly used, they offer the most critical evidence for decisions about acceleration. They also avoid some of the more intractable problems associated with attempts to define giftedness primarily in terms of anything that smacks of innate potential or capacity. Those who are not selected may rightly bristle at the suggestion that they are innately inferior. This is not as likely to occur when excellence is defined in terms of accomplishment.

Measures of developed reasoning abilities should be the second criteria for identifying children who are likely to profit from advanced instruction. Some students who show excellent reasoning abilities will be ready immediately for advanced instruction; some will be ready after a period of intensive instruction; and some will never be ready. Deciding which of these events is most probable requires clear thinking about the meaning of aptitude. Since defining the treatment is part of defining the aptitude, the first step is to identify the domains in which acceleration or advanced instruction is offered (or could be offered). In most school districts in the United States, the primary offerings consist of advanced instruction in areas such as mathematics, science, writing, literature, and (to a lesser extent) the arts. The best predictors of

success in any domain are typically achievement to date in that domain and the ability to reason in the symbol systems of that domain.

However, many students who have the potential for academic excellence do not meet the selection criteria of very high current achievement. The next goal, then, should be to find those students who do not currently exhibit academic excellence, but are most likely to develop it if given extra assistance. Because domain-relevant knowledge and skill are always important, one should look for students whose current achievement is strong, even though it is not stellar. Elsewhere (Lohman, 2004), I have used the example of selecting athletes for a college-level team. The best predictor of the ability to play basketball in college is the ability to play basketball in high school. Suppose, however, that the team could not recruit an athlete who had excelled at playing center in high school. One could cast a broader net and look at other athletes who had the requisite physical attributes (height, agility, coordination). But, it would be an extraordinary risk to recruit someone who had no experience in playing the game—or, better, who had demonstrated that he or she could not play the game. Rather, one would look for a player with at least moderate basketball playing skills in addition to the requisite physical attributes. Then, extra training would be provided.

A similar policy could guide the search for academically promising minority students. Suppose the criterion for admission to the gifted and talented program is scoring in the top 3% of the district in one or more domains of achievement or their associated reasoning abilities.[17] One could begin, for example, by identifying those students who score in the top 10% of the local distribution of achievement scores (but not in the top 3%). Among these high achievers, some will show higher-than-expected achievement the following year, whereas the majority will show somewhat lower achievement.[18] Those who show higher achievement the next year are more likely to show exceptionally strong reasoning abilities in the symbol system(s) used to communicate new knowledge. Therefore, among those with comparable achievement test scores, domain-critical reasoning abilities are the second most important aptitude. The reverse scenario holds for those who show high (but not stellar) scores on tests of reasoning abilities. For these students, domain achievement tests measure the second most important aptitude. In either case, within-group distributions of ability and achievement test scores can assist in the identification of the minority students most likely to develop academic excellence.

One reviewer of an earlier draft of this paper noted the similarities between these recommendations and the admission procedures used in many university-affiliated talent searches, such as The Center for Talented Youth at Johns Hopkins University, The Talent Identification Program at Duke University, The Center for Talent Development at Northwestern University, and The Rocky Mountain Talent Search at the University of Denver. Admission standards for entering the talent search require a 95–97 percentile rank score on an achievement test. Students are then administered verbal and quantitative reasoning tests, typically the SAT-I. Further, as recommended, students can qualify in either verbal or mathematical domains or both.

The policies discussed here differ, though, in the recommendation that school-based gifted and talented programs look not only at the students' ranks on achievement and ability tests when compared to all age or grade peers, but also at rank relative to those with similar backgrounds. Further, even though achievement should be given priority over reasoning abilities, it is generally wiser for school-based programs to test all students on both constructs, rather than screen on one test. If one test is used to screen applicants, then a much lower cutoff score should be used than is typically adopted (Hagen, 1980; see Lohman, 2003, for examples).

Student personality and affective characteristics also need to be taken into consideration, even though they typically show much stronger dependence on the particulars of the instructional environment than do the cognitive characteristics of achievement and reasoning abilities. Anxiety and impulsivity typically impede learning, especially when tasks are unstructured. Interest and persistence, on the other hand, typically enhance learning, especially in open-ended tasks (Corno et al., 2002). Identifications systems that take these sorts of characteristics into account will generally be more effective (Hagen, 1980).

Conclusions

1. *Except for very young children, academic giftedness should be defined primarily by evidence of academic accomplishment.* Measuring what students currently know and can do is no small matter, but it is much easier than measuring potential for future accomplishment. Good measures of achievement are critical. Start with an on-grade achievement test and, if necessary, supplement it with above-grade testing to estimate where instruction should begin. Look for other measures of accomplishment, especially production tasks that have well-validated rating criteria. For ELL students, attend particularly to mathematics achievement. High levels of current accomplishment should be a prerequisite for acceleration and advanced placement.

2. *Measure verbal, quantitative, and figural reasoning abilities in all students.* Keep in mind that the majority of students (minority and nonminority) show uneven profiles across these three domains and that the predictors of current and future achievement are the same in White, Black, Hispanic, and Asian American groups. Testing only those students who show high levels of achievement misses most of the students whose reasoning abilities are relatively stronger than their achievement. Because of regression of the mean, it also guarantees that most students will obtain lower scores on the ability test than they did on the achievement test (see Lohman, 2003).

3. *For young children and others who have had limited opportunities to learn, give greater emphasis to measures of reasoning abilities than measures of current accomplishment.* When accomplishments are few—either because students are young or because they have had few opportunities to learn—then evidence of the ability to profit from instruction is critical. As children mature and have opportunities to develop more significant levels of expertise, then measures of accomplishment should be given greater weight. In all cases, however, evidence of high levels of accomplishment trumps predictions of lesser accomplishments.

4. *Consider nonverbal/figural reasoning abilities as a helpful adjunct for both minority and nonminority admissions, but only as a measure of last resort.* Scores on the nonverbal tests must always be considered in conjunction with evidence on verbal and quantitative abilities and achievements. Defining academic giftedness by scores on a nonverbal reasoning test simply because the test can be administered to native speakers of English and ELL students serves neither group well. They are measures of last resort in identifying academically gifted children.

5. *Use identification tests that provide useful information for all students, not just the handful selected for inclusion in the gifted and talented program.* Teachers should see the test as potentially providing useful information about all of their students and how they can be assisted. A test that supposedly measures innate capabilities (or is easily misinterpreted as doing so) is rightly resented by those who are not anxious to see scores for low-scoring students entered on their records or used in ways that might limit their educational opportunities.

6. Learn how to interpret tables of prediction efficiencies for correlations. Even high correlations are much less precise for selection and prediction than most people think (see Lohman, 2003).

7. Clearly distinguish between the academic needs of students who show high levels of current accomplishment and those who show promise for developing academic excellence. Academic programs that combine students with high levels of current achievement with those who exhibit potential for high achievement often serve neither group well. When accelerating students, the primary criteria must be high levels of accomplishment in a domain. Common standards are necessary. Measures of developed reasoning abilities and other aptitudes (such as persistence) are best viewed as indicators of potential for future achievement. Students who show high potential, but only moderate levels of current achievement need different types of enrichment programs than those who currently show superior achievement.

8. Use common aptitude measures, but uncommon cutoff scores (e.g., rank within group) when identifying minority students most likely to profit from intensive instruction. Since the predictors of future accomplishment are the same for minority and White students, the same aptitudes need to be taken into account when identifying minority students who show the greatest potential for developing academic excellence. However, even with the best measures of aptitude, predictions will often be wrong. Because judgments about potential are inherently even more uncertain than judgments about accomplishment, a uniform cutoff score is difficult to defend when students come from different backgrounds. Both psychological and ethical issues must be addressed when making such decisions.

9. Do not confuse means and correlations. A biased selection procedure is one that, in any group, does not select the students most likely to profit from the treatment offered. Nonverbal reasoning tests reduce, but do not eliminate differences in mean scores between groups, and they are not the best way to identify those students who either currently exhibit or are most likely to achieve academic excellence. When used alone, they increase bias while appearing to reduce it. A more effective and fairer procedure for identifying academic potential is to look at both within-group and overall distributions of scores on tests that measure the most important aptitudes for successful learning in particular domains such as prior achievement and the ability to reason in the symbol systems used to communicate new knowledge in that domain.

References

Ackerman, P. L., & Kanfer, R. (1993). Integrating laboratory and field study for improving selection: Development of a battery for predicting air traffic controller success. *Journal of Applied Psychology, 78*, 413–432.

Achenbach, T. M. (1970). Standardization of a research instrument for identifying associative responding in children. *Developmental Psychology, 2*, 283–291.

Achter, J. A., Lubinski, D., & Benbow, C. P. (1996). Multipotentiality among intellectually gifted: "It was never there and already it's vanishing." *Journal of Counseling Psychology, 43*, 65–76.

Anastasi, A. (1980). Abilities and the measurement of achievement. In W. B. Schrader (Ed.), *New directions of testing and measurement; Measuring achievement: Progress over a decade* (pp. 1–10). San Francisco: Jossey Bass.

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

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89,* 369–406.

Assouline, S. G. (2003). Psychological and educational assessment of gifted children. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed, pp. 124–145). Boston: Allyn and Bacon.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* New York: Cambridge University Press.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8,* 205–238.

Bracken, B. A., & McCallum, R. A. (1998). *Universal nonverbal intelligence test*. Itasca, IL: Riverside.

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

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404–431.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83–118). Rowley, MA: Newbury House.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, England: Cambridge University Press.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

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: Erlbaum.

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

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 243–285). New York: Academic Press.

Forsyth, R. A., Ansley, T. N., Feldt, L. S., & Alnot, S. (2001). *Iowa test of educational development, Form A.* Itasca, IL: Riverside.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

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

Galton, F. (1972). *Hereditary genius*. Gloucester, MA: Peter Smith. (Original work published 1869)

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: BasicBooks.

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: Erlbaum.

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

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

Gustafsson, J. -E., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 186–242). New York: Simon & Schuster/Macmillan.

Hagen, E. P. (1980). *Identification of the gifted*. New York: Teachers College Press.

Hammill, D. D., Pearson, N. A., & Wiederholt, J. L. (1996). *Comprehensive test of nonverbal intelligence.* Austin, TX: PRO-ED.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms*. New York: Cambridge University Press.

Hedges, L. V., & Nowell, A. (1998). Black-White test score convergence since 1965. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 149–181). Washington, DC: Brookings Institution Press.

Hoover, H. D., Hieronymous, A. N., Frisbie, D. A., & Dunbar, S. B. (1993). *Iowa test of basic skills, Form K: Survey battery.* Chicago: Riverside.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). *The Iowa test of basic skills, Form A.* Itasca, IL: Riverside.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 267–300). New York: Wiley.

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., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership: Exemplified by the role of spatial visualization for becoming an engineer, physical scientist, or artist. *Journal of Applied Psychology, 78,* 250–261.

Irving, S. H. (1983). Testing in Africa and America. In S. H. Irvine & J. W. Berry (Eds.), *Human assessment and cultural factors* (pp. 45–58). New York: Plenum Press.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

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.

Keith, T. Z., & Witta, E. L. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC-III: What does it measure? *School Psychology Quarterly, 12,* 89–107.

Koretz, D., Lynch, P. S., & Lynch, C. A. (2000). The impact of score differences on the admission of minority students: An illustration. *Statements of the National Board on Educational Testing and Public Policy, 1*(5). Retrieved February 2, 2005, from http://www.bc.edu/research/ nbetpp/publications/v1n5.html

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14,* 389–433.

Laboratory of Comparative Human Cognition. (1982). Culture and intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 642–722). New York: Cambridge University Press.

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, 95,* 482–494.

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

Lohman, D. F. (1996). Spatial ability and *g*. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and assessment* (pp. 97–116). Hillsdale, NJ: Erlbaum.

Lohman, D. F. (2000). Complex information processing and intelligence. In R.J. Sternberg (Ed.), *Handbook of human intelligence* (2nd ed., pp. 285–340). Cambridge, MA: Cambridge University Press.

Lohman, D. F. (2003). *Tables of prediction efficiencies*. Retrieved February 2, 2005, from http://faculty.education.uiowa. edu/dlohman

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. 41–55). New York: Routledge/Falmer.

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

Lohman, D. F. (in press). An aptitude perspective on talent: Implications for identification of academically gifted minority students. *Journal for the Education of the Gifted.*

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., & Hagen, E. P. (2001c). *Cognitive Abilities Test (Form 6): Interpretive guide for teachers and counselors.* Itasca, IL: Riverside.

Lohman, D. F., & Hagen, E. P. (2002). *Cognitive Abilities Test (Form 6): Research handbook.* Itasca, IL: Riverside.

Lubinski, D. (2003). Exceptional spatial abilities. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 521–532). Boston: Allyn and Bacon.

Lubinski, D., & Benbow, C. P. (1995). An opportunity for empiricism [Review of *Multiple intelligences*]. *Contemporary Psychology, 40,* 935–937.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7,* 107–128.

Martinez, M. (2000) *Education as the cultivation of intelligence.* Mahwah, NJ: Erlbaum.

Miller, J. G. (1997). A cultural-psychology perspective on intelligence. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Intelligence, heredity, and environment* (pp. 269–302). New York: Cambridge University Press.

Mills, C., & Tissot, S. L. (1995). Identifying academic potential in students from under-represented populations: Is using the Ravens Progressive Matrices a good idea? *Gifted Child Quarterly, 39,* 209–217.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12,* 252–284.

Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test: Multilevel technical manual.* San Antonio, TX: Harcourt Brace.

Naglieri, J. A., & Das, J. P. (1997). *Cognitive assessment system.* Itasca, IL: Riverside.

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. (2005). Increasing minority children's participation in gifted classes using the NNAT: A response to Lohman. *Gifted Child Quarterly, 49,* 29–36.

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

Oakland, T., & Lane, H. B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing, 4,* 239–252.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart, and Winston.

Piaget, J. (1952). *The origins of intelligence in children.* New York: International Universities Press.

Plomin, R., & De Fries, J. C. (1998). The genetics of cognitive abilities and disabilities. *Scientific American, 278,* 62–69.

Raven, J. (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.

Richert, E. S. (2003). Excellence with justice in identification and programming. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 146–158). Boston: Pearson Education.

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.

Scarr, S. (1981). *Race, social class, and individual differences in IQ.* Hillsdale, NJ: Erlbaum.

Scarr, S. (1994). Culture-fair and culture-free tests. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 322–328). New York: Macmillan.

Shepard, R. N. (1978). Externalization of mental images and the act of creation. In B. Randhawa & W. Coffman (Eds.), *Visual learning, thinking, and communication* (pp. 133–190). New York: Academic Press.

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 76,* 347–376.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp .263–332). New York: Macmillan.

Snow, R. E., & Yalow, E. (1982). Education and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 493–585). New York: Cambridge University Press.

Sternberg, R. J. (1982). Reasoning, problem solving, and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 225–307). New York: Cambridge University Press.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* Cambridge, England: Cambridge University Press.

Sternberg, R. J., & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Development, 51,* 27–38.

Thorndike, R. L., & Hagen, E. (1993). *The cognitive abilities test: Form 5.* Itasca, IL: Riverside.

Thorndike, R. L., & Hagen, E. (1996). *The Cognitive Abilities Test (Form 5): Interpretive guide for school administrators.* Itasca, IL: Riverside.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs,* 1.

U. S. Department of Education, Office of Education Research and Improvement. (1993). *National excellence: A case for developing American talent*. Washington, DC: U.S. Government Printing Office.

Author Note

End Notes

1. A good example of the disconnect between stimuli and how they are processed is provided by a test of spatial ability that I devised to help select air traffic controllers. I called it the Verbal Test of Spatial Ability. Items were presented verbally (e.g., "Imagine that you are walking north. You turn right at the first corner, walk one block, and then turn left. In what direction are you facing?"), and required a verbal response ("North"). However, the test was one of the best measures of spatial ability in the selection battery (Ackerman & Kanfer, 1993).

2. These sorts of confusions are typically not detected by statistical tests for item bias. This is because strategies such as picking an associate of the last term in the problem stem will often lead to the keyed answer on simple problems. The student thus appears to be able to solve some easy items, but not the more difficult ones. Bias statistics look for cases in which an item is consistently harder or easier for students in a particular group, given the number of problems the student answers correctly.

3. I agree with Lubinski (2003) that we should measure spatial abilities routinely in talent searches, especially if we can provide instruction that capitalizes on these abilities and furthers their development. However, one should measure these abilities explicitly rather than inadvertently.

4. Children's storybooks provide interesting examples of this variation in styles of depicting people, animals, and objects in different cultures. The variations are reminiscent of cultural variations in the onomatopoeic words for animal sounds.

5. For an excellent summary of the role of parents in developing the deductive reasoning abilities of their children, see J. Raven (2000). Raven argues that such development is promoted if parents involve children in

their own thought processes. Such parents are more likely to respect their children and "to initiate a cyclical process in which they discover just how competent their children really are and, as a result, become more willing to place them in situations that call for high-level competencies" (p. 33).

6. The test score variance that is explained by the factor is given by the square of the tests loading on the factor. For example, if a test loads .6 on the *Gf* factor, then the *Gf* factor accounts for 36% of the variance on the test.

7. Although Jensen (1998) disagrees, a much longer list of other notables agrees (Cronbach, 1990; Horn, 1985; Humphreys, 1981; Plomin & De Fries, 1998). Indeed, Humphreys was fond of pointing out that, in the Project Talent data, heritability coefficients were as high for a test of knowledge of the Bible as for measures of fluid reasoning ability.

8. Since Thurstone (1938), test developers have constructed tests that measure different abilities by increasing the representation of tests (and items) that have lower loadings on g and higher loadings on group factors. CogAT test batteries were not constructed in this way. Each battery is designed to maximize the amount of abstract reasoning that is required and is separately scaled. Correlations among tests were computed only after the test was standardized.

9. To be significant, the difference must be at least 10 points on the Standard Age Score scale (mean = 100, *SD* = 16), and the confidence intervals for the two scores must not overlap. The confidence intervals will be wide if a student responds inconsistently to items or subtests in the battery.

10. Standard Age Scores (SAS) have a mean of 100 and a standard deviation of 16.

11. For frequencies of different score profiles in the full population, see page 86 in Lohman and Hagen (2001c). For frequencies of score profiles for stanine 9 students, see page 125 in Lohman and Hagen (2002).

12. For the multilevel battery, the KR–20 reliabilities average .95, .94, and .95 for the Verbal, Quantitative, and Nonverbal batteries, respectively (Lohman & Hagen, 2002).

13. This statistical fact of life is commonly overlooked by those who would insist on high scores in all three content domains on CogAT to qualify for inclusion in G&T programs. It is also why the CogAT authors recommend that schools not use the Composite score for this purpose.

14. In these analyses, we predicted achievement from a composite score that averaged across the three CogAT

batteries (which was entered first into the regression) and then the scores of the three CogAT batteries (which were entered simultaneously in the second block). The nonverbal score typically had a negative regression weight, which was larger for minority students than for White students.

15. Although one can compute the unstandardized regression coefficients for with-sex regressions from the *b* weights reported in Tables 2 and 3, standardized regression coefficients (β's) required additional, within-sex analyses. These analyses are not reported here.

16. See Lohman (2003) for tables that display these prediction efficiencies for different levels of correlation between tests.

17. The CogAT authors have consistently argued that such decisions should not be made on the basis of the CogAT composite score (e.g., Hagen, 1980; Lohman & Hagen, 2001b, p. 127; Thorndike & Hagen, 1996, p. 159). Academic excellence should also be identified within the major curricular domains. Some students will excel in multiple domains; most will not.

18. Regression to the mean is most evident in scores that report rank within group (such as percentile ranks) and least evident on attainment scores (such as developmental scale scores on achievement or ability test batteries).