Chapter 49

Identifying Academically Talented Students: Some General Principles, Two Specific Procedures

David F. Lohman

Abstract Talent signifies potential for the attainment of excellence in some domain. Procedures for identifying academic talent that were developed at the turn of the twentieth century emphasized the measurement of a cognitive potential that was thought to forecast successful learning in any domain or learning context. While not rejecting the utility of the measures of cognitive ability devised to estimate this learning potential, most psychologists now see talent as a more variegated concept and the development of expertise as dependent on much more than superior cognitive ability. The concept of aptitude offers a useful way to integrate these disparate strands into a coherent theory. In this chapter, I discuss the identification of academically talented children from the perspective of aptitude theory. Aptitude refers to the degree of readiness to learn and to perform well in a particular situation or domain. The primary aptitudes for academic success are (1) prior knowledge and skill in a domain, (2) the ability to reason in the symbol systems used to communicate new knowledge in that domain, (3) interest in the domain, and (4) persistence in the type of learning environments offered for the attainment of expertise in the domain. Although the principles discussed here are useful for all students, they are particularly important for the identification of academically promising minority students. The chapter concludes with examples of two procedures for combining ability test scores, achievement test scores, and teacher ratings in a principled way to assist in the identification of a talent pool.

D.F. Lohman (✉)
The University of Iowa, ~~Ames,~~ IA, USA
e-mail: david-lohman@uiowa.edu

Keywords Talent · Aptitude · Expertise · Ability · Group ability tests · Achievement tests · Regression to the mean · Teacher ratings

Introduction

The purpose of this chapter is to justify and then illustrate a way to conceptualize the identification of talent that acknowledges, but goes well beyond the concepts of giftedness first advanced in the waning years of the nineteenth century. Although talent can be in any area, the focus here is on academic talent. The approach is grounded in over 80 years of research on the conceptualization and measurement of aptitude (see Corno, Cronbach et al., 2002). The chapter is based on several other papers that I have recently published on this topic, most notably a monograph issued by the National Research Center on the Gifted and Talented (Lohman, 2005a). Most of these papers use data from Form 6 of the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001), which I now coauthor with Elizabeth Hagen.[1] Those who desire elaboration of points made here are encouraged to look at the original documents for less abbreviated discussions of the issues. Those familiar with this work may want to skip directly to the final section. There I describe two procedures for developing talent identification systems

[1] I include this information to disclose at the outset any conflict of interest. It also explains why I have used CogAT and ITBS data to illustrate points that I make. However, on more than one occasion I found it necessary to explain how CogAT differs from other group ability tests. Although some will interpret this as self-promotion, my real concern was to correct the widespread misconception that all such tests are more or less exchangeable.

L.V. Shavinina (ed.), *International Handbook on Giftedness*,
DOI 10.1007/978-1-4020-6162-2_49,

that try to incorporate some of the principles that I advocate.

The identification of academically gifted children depends not only on a clear definition of giftedness but also on the kind of educational programs that will be made available to children who are selected to participate in them. Programs range from highly selective, special schools for profoundly gifted children to enrichment activities that are made available to all children in a school who show interest in a particular activity and some talent for it. Clearly the specific procedures that one would use to identify those children who might succeed in a highly selective school for the gifted would not be appropriate for a program that follows a School-Wide Enrichment Model (Renzulli, 2005). Nevertheless, the same principles govern each. It is these principles that are the focus of this chapter.

Modern conceptions of ability—while not disavowing the practical import of ability tests—interpret the psychological constructs that they measure quite differently than their developers did. In particular, abilities are seen as forms of emerging expertise that are developed through participation in and internalization of activities valued by the culture(s) in which the child lives (see, e.g., Horn & Blankson, 2005). Talent is envisioned as a readiness or propensity to acquire particular kinds of expertise. Talents of greatest interest will be those needed to develop those forms of expertise valued by a society through the formal training systems available to the individual in that society. Academic talent is thus critical for all of those forms of expertise for which formal education is required or for which it serves as the gatekeeper.

An increasingly important problem for all talent identification systems is to find better ways to identify academically talented poor and minority students. Attempts to increase the representation of these students, however, have been made difficult by recurring misconceptions about the nature of academic giftedness, the interpretation of measures of ability and achievement commonly used to identify gifted students, and the kinds of the educational programs that have been developed to serve gifted students. Many of the procedures that are widely used to measure academic talent are not effective when children have had markedly different opportunities to develop those talents. Attempts to remediate this problem without first understanding the psychological foundations of talent are at best only partially effective and, more commonly, actually miss the most academically talented minority students. Assuming that nonverbal ability tests such as the ~~Normal~~ Battery of CogAT will do the trick is the most glaring example of this failure (Lohman, 2005b). Therefore, I first discuss misconceptions about giftedness that have thwarted efforts to identify academically talented minority students. Chief among these is the fallacy that intelligence tests can be constructed that measure innate ability. I then present an alternative model for identifying academically talented students that is grounded in modern theories of aptitude.

In brief, my argument is that (a) academic talent is best understood as aptitude for the kinds of expertise that can be developed through schooling; that (b) the primary aptitudes for academic learning are current knowledge and skill in a domain, the ability to reason in the symbol systems used to communicate new knowledge in the domain, interest in the domain, and persistence in the pursuit of excellence; and that (c) inferences about academic talent are most defensible when made by comparing a student's behavior to the behavior of other students who have had similar opportunities to acquire the knowledge and skills measured by the aptitude tests; however, (d) educational programming and placement should be based primarily on evidence of current accomplishments compared to all other students.

Historically, the identification of giftedness focused exclusively on cognitive competence as estimated by an IQ test. This is still an important theme in some quarters—especially among those whose primary concern is the identification of profoundly gifted students. However, even here it is not apparent that the effort spent in deciding whether a child's IQ is 158 instead of 145 is as useful as a more careful documentation of the child's academic development (and interests) in different domains of study. At the end of the day, the real problem is how best to intervene to assist children in developing their abilities and talents.

Although it does not dismiss the notion of general ability (except perhaps in the work of Ericsson, Nandagopal, & Roring, 2005), modern research on the development of competence starts not with general ability but rather with the kind of expertise that one hopes to forecast. Deciding which students are most likely to develop that type of expertise requires an understanding of what constitutes expertise in the domain and how it might be developed. Typically, different combinations of personal characteristics are needed for success at different stages in the development of expertise. For example, success in the early reading or mathematics requires somewhat different abilities than success in later literacy or advanced mathematics. Indeed, precocious reading does not forecast general academic excellence (Jackson, 1988). Next one looks at the demands and opportunities of the different educational paths offered for those who wish to develop this kind of expertise. Students must always develop their talents within the educational programs that are available to them. Even the best of these programs will better serve some children than others. What must students know and be able to do to succeed in alternative routes to the attainment of expertise? The student who might have difficulty learning alone (e.g., computer-assisted construction) might succeed more readily when learning with others (e.g., a symposia or discussion group). Always one should strive for "congruence between the criteria used in the identification process and the goals and types of services that constitute the day-to-day activities that students will pursue" (Renzulli, 2005, p. 11).

Defining Giftedness

How best to identify gifted children is one of the most persistent and controversial topics in the field of gifted education (VanTassel-Baska, 2000). Much of the controversy stems from different beliefs about the meaning of the term *gifted*. Should giftedness be restricted to academic domains, or should it include artistic, athletic, leadership, and other types of competence valued by society? In the academic domain, should it be based on evidence of superior academic accomplishments or on a measure of academic potential, such as IQ?

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

Even a cursory consideration of these questions reveals that we are not interested in ability for ability's sake, but rather in ability *for* something. Most commonly the goal is to identify those children who either currently display or are likely to develop excellence in those skills valued by society and that are therefore taught in its schools. Identifying such students is a more manageable problem than trying to measure the hundreds of ways in which people differ and then creating programs that are uniquely tailored to each kind of exceptionality. Put differently, those who take an ability-centered approach to the identification of giftedness have no basis other than parsimony for designating one ability as more important than any other ability. For example, it is only when we add the criterion of utility that general crystallized abilities become much more important than general spatial or general memory abilities in the identification of academic giftedness. Crystallized abilities better predict school achievement for all children even though general crystallized, fluid, spatial, and memory abilities have equal stature in the modern theories of human abilities.

Giftedness as Relative to the Norm Group

Judgments of exceptionality depend on the norm group. Scores that are unusual in one cohort often are not unusual in another. There are many examples of studies that demonstrate how differences in norms can confound efforts to identify gifted students. For example, Shaunessy, Karnes, and Cobb (2004) administered the Culture Fair Intelligence Test (CFIT; Cattell & Cattell, 1965), the Standard Progressive Matrices (Raven; Raven et al., 1983), and the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997) to 196 predominantly Black students in a poor rural school district. Their goal was to see which test identified the most gifted students. To do this, they compared the number of students who fell in 5-point percentile bands on each test beginning at the 80th percentile.

The NNAT, which has the most recent norms, identified only three students as falling above the 80th age percentile rank (PR); the Progressive Matrices Test, which has some normative data collected in the 1980s, identified 18 students; and the CFIT, which has the oldest and least defensible norms, identified 36 students.[2] A more recent study that compared the performance of Hispanic English language learners (ELL) in kindergarten through grade 6 with the performance of non-ELL students also showed the inadequacy of the outdated norms for the Raven. In all, approximately 1200 students were administered the Standard Progressive Matrices, the NNAT, and Form 6 of the Cognitive Abilities Test in a counterbalanced order. Across all grades, the mean SAS score on CogAT Nonverbal battery and the mean NAI score on NNAT were the same for non-ELL students (100.7) and similar for ELL students (92.4 on CogAT and 90.7 on NNAT). However, when converted to a similar IQ-like metric, the corresponding mean scores for Raven were 111 and ~~108~~. Because the Raven norms were too easy, the test vastly over-identified the number of "gifted" children, especially in the early grades.[3] Surprisingly, the NNAT also substantially over-identified the number of high-scoring children because of errors in norming the test.

Even if tests are normed on the same population, the common practice of accepting the highest score across several different tests that measure the same ability is fundamentally misguided. It assumes that the highest score in a set of different test scores best estimates a student's ability. This is not true. The best estimate is usually given by the average of these scores. The highest score among a set of ostensibly parallel assessments is actually one of the most error-laden scores, and thus most likely to regress to the mean.

Flynn Effect

National norms for both ability and achievement tests have been changing for as long as we have been norming tests. Scores on ability tests have been rising at the rate of about three IQ points per decade since the 1920s. This increase is sometimes called the *Flynn effect* after the researcher who first systematically documented it in many countries (see, e.g., Flynn, 1987). Growth has been even larger on nonverbal tests such as the Progressive Matrices and was unabated in the most recent studies (see Raven, 2000). This is one reason that major ability and achievement tests are re-normed every 5–10 years. Even when two tests are normed in the same year, however, samples of examinees differ, and so norms for the two tests are not necessarily the same. Therefore, a selection rule that defines admission in terms of IQ or national percentile rank will not admit the same number of students in different years or when different tests are used, especially if the norms are not of equal recency and quality.

Demographic Changes in the Population

Norms also change to reflect demographic shifts in the population. For example, norms for verbal tests change as the number of the bilingual children and adults increases. Historically, these changes have been relatively small. However, recent increases in the number of bilingual children and adults will result in a relative "softening" of the norms on these tests. This will lead to a relative increase in the number of high-scoring non-ELL children.

Norms on achievement tests, on the other hand, are based entirely on the fraction of the population attending school. Because low-achieving students are more likely to drop out before completing high school, a given percentile rank (say 95) means better performance compared to all children in the population for the 12th-grade student than for the 3rd-grade student.

The Importance of Local Norms

The Flynn effect demonstrates that judgments about exceptionality depend on the national norms that are

---

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

[3] Although overall mean scores across grades did not differ for CogAT and NNAT, the variability of scores was much greater on NNAT than on CogAT or Raven. Other analyses of the NNAT standardization data show the same elevated SD in the Nonverbal Ability Index (NAI), with the population of SD exceeding 21 at level A. Failure to set the SD of NAI scores to 15 results in approximately three times as many children receiving scores greater than 130 at the primary level than would be observed had the SD been properly set to 15 (see Lohman, Korb, & Lakin ~~in press~~)

used to interpret scores. More students will obtain high scores when older norms are used than when more recent norms are used. However, differences between schools in the same state are many times greater than differences between cohorts of students in different decades. This is important because the need for special services depends not so much on a student's standing relative to age- or grade mates nationally, but on the student's standing relative to the other students in the class. Talent searches and district-wide programs that recruit students from different schools need the common standard of national norms. National norms also provide important information on a student's relative standing on the different abilities measured by the test. Individual schools, however, rarely replicate the nation in their distribution of ability or achievement. In about 5% of the schools in the nation, the *average* student scores at the 95th percentile on the Iowa Test of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001). Surely the students in these classes are quite capable. But it is unlikely that a student who scores at the 98th national percentile in such a class will be as mismatched with the common curriculum as the student who scores at the 98th national percentile in a class in which the typical student scores at the 50th percentile on the ITBS. In short, although both national and local norms have important uses, decisions about identification and acceleration are often best made using local norms. Many publishers offer local norms when the school or district tests all children in a particular grade.[4]

Age Norms

Judgments about academic potential often assume a different cohort than either a national or a local grade cohort. Suppose that we discover that the student whose achievement is exceptional is actually a year older than the other students in the class. Although instruction should be geared to the child's achievement, would one still consider the child "gifted"? Conversely, suppose a child is considerably younger than her classmates or has attended school irregularly. Should her scores be compared with others in the same grade when estimating her ability to learn?

One of the primary differences between ability and achievement tests is that ability tests report scores relative to age mates.[5] Ability is an inference about rate of learning given equal opportunity to learn. We use age as a yardstick in measuring ability because it is a useful surrogate for "total amount of experience in the culture." If the abilities are those that can be developed in the course of everyday interactions with the culture, then comparisons to one's age cohort provide important information. If the abilities can be developed through school experiences, then comparisons with those who have had similar amounts of education (i.e., grade norms) are also helpful. However, even small differences in the choice of age cohort (e.g., 6 years 0 months versus 6 years 10 months) can make a large difference in whether a particular score is considered exceptional. If the child was ill for several months or lived in the culture for only half of her life, would norms based on her age cohort be most appropriate for inferences about her ability to learn?

Subgroup Norms

For those whose experiences differ markedly from the norm, aptitudes need to be judged relative to a different cohort. *Always, the preferred comparison group would be those who have had roughly similar opportunities to acquire the abilities sampled by the test.* Concretely, one should look at the performance of the ELL child relative to other ELL children who have had roughly similar amounts of exposure to English. The "fair" procedure of comparing all to the same age or grade group regardless of their experience is equivalent to the "fair" procedure of comparing all children to a common score distribution, regardless of age. Furthermore, as shown elsewhere (Lohman, 2005a) one need not develop rigorous norms tables or compare the child only to the handful of others who have had similar experiences. Rough classifications (such as ELL versus native speakers) go a long way to correcting the problem.

[4] Procedures described in the final section of this chapter show how to approximate these norms.

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

Assessments in Other Languages

Should a test also be administered in the child's other language(s)? If the goal is to assess the full extent of a child's cognitive competence, then the test(s) needs to be aligned not only to the language(s) but also to the culture(s) of the child. This is more commonly a problem when making judgments about cognitive impairments than about readiness to profit from advanced or more rapidly paced instruction. In either case, oral language can introduce construct-irrelevant variance into the testing situation. One can reduce this impact and still make some inference about verbal abilities by presenting problems that require verbal processing, but that use only pictures. This allows the child to use any language when attempting items. However, such tests do not measure verbal reasoning very well. Many concepts—especially those that require nuanced judgments about meaning—cannot be displayed in pictures. Therefore, such tests substantially underrepresent those aspects of reasoning most clearly captured by reasoning tests that ~~can~~ use words.

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

Scaling Effects

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

Summary

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

Is Giftedness Developed?

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

Three facts are commonly overlooked in discussions of the extent to which genetic factors explain or account for individual differences in intelligence. First, the contribution of heredity is typically as high for measures of achievement as it is for measures of ability. The distinction between ability and achievement tests cannot to be made on the basis of the contribution of heredity to individual differences in the scores (Cronbach, 1976). Second, estimates of heritability vary substantially across cohorts of people who share a common genetic heritage. For example, Turkheimer, Haley, Waldron, D'Onofrio, and Gottesman (2003) investigated how the heritability for IQ scores on the Wechsler Intelligence Scale for Children (WISC) at age seven varied as a function of socioeconomic status (SES) for 319 twins. The best estimate of heritability for the lowest SES children was less than .20. This means that individual differences in the WISC IQ scores of these twins were almost entirely due to variations in their pre- and post-natal environments. For the highest SES children, on the other hand, the estimated heritability was at least 80% of the variance. Therefore, the contribution of the environment—both that portion shared by both children and that portion unique to each child—was vastly more important for low-SES children. Third, even a high heritability does not imply immutability. Individuals can show marked changes in the trait over time and their ranks within the group change substantially while the heritability remains the same. Indeed, this is the case for those abilities that we measure with intelligence tests.

Giftedness as a Category Label

The act of naming something enables communication with others. But it can also reinforce the perverse human tendency to misrepresent as categorical a characteristic that varies continuously. We speak of *learning disabled* or *gifted* students as if these labels represented discrete categories rather than arbitrary portions of continuously varying score distributions. If two well-respected tests give different ability scores for the same child, one saying that the child is gifted and the other reporting a lower score, many would dismiss one outcome—usually the lower score. But such disagreement between tests is the rule, not the exception. Suppose we define *gifted* as scoring in the top 3% of the distribution of intelligence. We administer two of the best intelligence tests—the Stanford-Binet V and the WISC-IV—to each student. Full-scale IQ scores on these tests correlate $r = .84$ (Roid, 2003). This means, however, that only about half of the students who score in the top 3% on one test will also score in the top 3% on the second test (Lohman & Korb, 2006). If the interval between test administrations is longer than a few weeks, then even fewer would merit the label *gifted* on both tests.

This is not what most test users expect, in part, because when we think about gifted children, we tend to envision those children who most clearly exemplify the category. These will generally be those children with the most extreme scores. And although the scores for these children also differ across tests, both scores are likely to fall above the cut score. For other students, we have an unfailing tendency to focus on those scores that are consistent across occasions because they confirm our expectations of consistency. For example, children whose performance is exceptional at age 6 but not at age 10 will often be dismissed as "not really gifted," even though their performance at age 6 truly was exceptional.

Confirmation bias is widespread in human reasoning (Nickerson, 1998). However, no matter where we set the cut score on the upper tail of the score distribution, many more students will be near that cut score than far above it. When the scores of these students regress toward the mean on retest many will fall below the cut, and the scores of an equally large group of students who previously failed to make the cut will now rise above it. Regression to the mean is inevitable whenever the two sets of scores are not perfectly correlated.[6] Indeed, it is another way of

[6] Technically, this holds only if the variability of scores is the same for the two sets of scores. This will always be the case when using IQ scores, percentile ranks, and other status scores.

saying that two variables are not perfectly correlated. The lower the correlation or the more extreme the initial score, the greater the regression. Keeping track of everyone—not just the cases that stand out clearly in our minds—helps us combat the confirmation and typological biases in our thinking.

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

The critical mistake here is to assume that ability is fixed, not constantly developing. It ignores the fact that to maintain a particular rank (e.g., IQ), a child must not only get better each year but must improve at the same rate as others who had the same initial score. Using status scores such as percentile ranks (or derivatives such as IQs) masks this year-to-year growth. If the same dimension were labeled "language development" rather than "giftedness," then we would expect to find some children whose development was unusual at one point in time but not unusual at a later point in time. Reading is unusual for a 3-year-old. It is not unusual for a 6-year-old. However, as is the case with many gifted behaviors in young children, precocious readers often have IQ's in the moderately above-average range. Indeed, "the literature provides little support for the assumption that giftedness in childhood is an enduring and unchanging property of the individual. Rather, different forms of giftedness emerge at different ages in different children" (Jackson, 2003, p. 470). Those who identify gifted students can thus inadvertently become gatekeepers for the precocious rather than advocates of developmentally appropriate instruction for all students.

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

An Aptitude Theory of Academic Talent

An aptitude approach to understanding academic talent is very much concerned with abilities but in a different way than in most theories of giftedness. The focus is on *all* of the aptitudes that must be brought to bear to accomplish something. In particular, the goals are to identify either those children who currently display academic excellence and are likely to continue to display it or those children who show less exceptional levels of accomplishment but are likely to develop it if given special assistance. The first point, then, is that academic giftedness is best understood in terms of aptitude to acquire the knowledge and skills taught in schools that lead to forms of expertise that are valued by society. We are interested in ability tests only because they help identify those who may someday become excellent engineers, scientists, writers, etc. In other words, we are interested in abilities because they are indicants of aptitude. They are not the only indicants but one important class of indicants.

A Definition of Aptitude

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

So, what exactly do we mean by *aptitude*? Although often rooted in biological predispositions, it is not something that is fixed at birth. School achievements commonly function as aptitudes—for example, reading skills are important aptitudes for school learning. Indeed, aptitude encompasses much more than cognitive constructs such as ability or achievement. Persistence is an important aptitude in the attainment of expertise. Often, particular ability, interest, and motivational characteristics work together and thus are best understood as an aptitude complex (Ackerman, 2003; Lubinski & Benbow, 2000). For example, the effects of general ability are moderated by anxiety: Although high-ability students generally learn best in less structured environments, those who are highly anxious typically do better in more structured environments. Finally, and most importantly, the term *aptitude* is not a descriptor of a person that is somehow independent of context or circumstance. Indeed, *defining the situation or context is part of defining the aptitude*. Changing the context changes in small or large measure the personal characteristics that influence success in that context. Aptitude is inextricably linked to context.

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

Effects of Context

Understanding which characteristics of individuals are likely to function as aptitudes begins with an understanding of the demands and affordances of target tasks and the contexts in which they must be performed. This is what we mean when we say that defining the situation is part of defining the aptitude (Snow & Lohman, 1984). The affordances of an environment are what it offers, makes likely, or makes useful. Discovery learning often affords the use of reasoning abilities; direct instruction often reduces the need for these abilities. Unless we define the context clearly, we are left with distal measures that capture only some of the aptitudes needed for success. This is why *g*-like measures of ability correlate imperfectly with success in any particular school task, especially when students are allowed a choice over what they study and how they might go about it. On the other hand, averaging across learning situations and outcome measures obscures the impact of the particular abilities and magnifies the relative importance of *g*.

Physical skills are similar. General physical fitness is the best (albeit typically weak) predictor of success across a wide range of athletic skills. However, more specific physical abilities and skills become more important aptitudes as the range of athletic events narrows. Finally, when it is available, current skill in a sport is generally the best predictor of performance in the immediate future.

Inferring Aptitudes

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

In the second way of inferring aptitude, we attempt to identify other tasks that require similar cognitive or affective processes and measure the individual's facility on those tasks (Carroll, 1974). Because these measures only predict success on the task, they will more often error in identifying those students who will excel in learning the task itself. Even if by some miracle one could carefully specify all of the cognitive aptitudes for learning, one would also need to specify the needed affective and conative aptitudes as well. One of the reasons why measures of current accomplishment in a domain are important measures of aptitude for future learning is that they already incorporate this information.

Scholastic Aptitudes

The two most important aptitudes for academic learning are reasoning abilities and knowledge and skill in the domain of instruction. The relative importance of prior achievement and reasoning abilities will depend on the demands and affordances of the instructional environment and on the age and experience of the learner. In general, prior achievement is more important when new learning is like the learning sampled on the achievement test. This is commonly the case when the interval between old and new learning is short. With longer time intervals between tests or when content changes abruptly (as from arithmetic to algebra), then reasoning abilities become more important (Lohman & Korb, 2006). Novices typically rely more on knowledge-lean reasoning abilities than do domain experts. Because children are universal novices, their reasoning abilities are more important in the identification of academic talent, whereas evidence of domain-specific accomplishments is relatively more important for adolescents (Hagen, 1980). Therefore, if the goal is to identify those students who are most likely to show high levels of future achievement, both current achievement and domain-specific reasoning abilities need to be measured. Analyses of the CogAT-ITBS data (Lohman & Korb, 2006) suggest that the two should be weighted approximately equally.

Measures of Domain Knowledge and Skills

A critical requirement of most academic tasks is domain knowledge and skill (Glaser, 1992). When they are available, measures of prior knowledge and skill are therefore usually the best predictors of success in academic environments, especially when new learning depends heavily on old learning. Measures of current knowledge and skill include on-grade-level and above-grade-level achievement tests and well-validated performance assessments such as rankings in debate contests, art exhibitions, and science fairs. The sorts of group-administered achievement tests that are commonly used in schools provide useful information about students' general reading achievement and mathematics skills. However, by design such tests present only the most general concepts in science, social studies, and other particular domains of study. To better understand students' development in these domains, schools might investigate end of course examinations in particular subjects—those exams used either in local schools or available from major test publishers. Inventories of conceptual and factual knowledge in a domain can also provide critical information on this aspect of academic development. These are frequently overlooked, often because there is no easy way to rank all children on the same dimension. Most achievement tests—especially those designed for elementary school children—contain relatively little content knowledge. However, studies of the development of expertise show that to develop competence in an area of inquiry, students must construct rich networks of well-organized factual and conceptual knowledge (Bransford, Brown, & Cocking, 2000). This knowledge can be quite localized, especially when learning is self-directed. Bright children assemble vast amounts of knowledge about specific topics that are at best represented only superficially on achievement tests. An achievement test that is designed to be fair to all children can hardly be expected to reveal much about the specialized knowledge a student has acquired. In this respect, the dilemma that confronts those who would assess gifted children is the same dilemma that has stymied those who investigate adult intelligence. Hunt (2000) suggests that we might do a better job if the metaphor that guided the construction of the assessment were to conduct an inventory rather than a survey.

Short-term educational decisions should therefore rely primarily on evidence of current accomplishment in a domain. Other aptitudes enter the picture, though, with each step one takes into the future. For example, given the same type of instruction, continued improvement in a domain requires interest or at least dogged persistence. More commonly, continued success requires a new mix of abilities: Algebra requires some skills not needed in arithmetic; critical reading requires skills not needed in beginning reading. Teachers, teaching methods, and classroom dynamics also change over time, each requiring, eliciting, or affording the use of somewhat different personal characteristics. Indeed, in most disciplines, the development of expertise requires mastery of new and, in some cases, qualitatively different skills at different stages. Sometimes the critical factor is not only what is required for success but also what is allowed or elicited by the new context that might create a stumbling block for the student. For example, in moving from a structured to a less structured environment, a student may flounder because he is anxious or is unable to schedule his time. Indeed, the attainment of expertise often has as much to do with inaptitudes as aptitudes.

Measures of Fluid Reasoning Abilities

The second most important set of personal characteristics for academic learning are the ability to go beyond the information given; to make inferences and deductions; and to see patterns, rules, and instances of the familiar in the unfamiliar. The ability to reason well in the symbol system(s) used to communicate new knowledge is critical for success in learning. Academic learning relies heavily on reasoning (a) with words and about the concepts that they signify and (b) with quantitative symbols and the concepts that they signify. Thus, the critical reasoning abilities for all students (minority and majority) are verbal and quantitative. This is not a new idea. It has been known since aptitude tests were first devised to predict success in school and college. Figural reasoning abilities are less important and show lower correlations with school achievement (Lohman, 2005b).

Individual versus group ability tests. Group ability tests are commonly viewed as rough screening measures that will at best give a global and somewhat labile estimate of a child's abilities. Individually administered ability tests are generally considered the gold standard. As with most stereotypes, this is not always the case. The generalization is most likely to hold when the group-administered test is relatively short or samples only a portion of the cognitive domain. For example, nonverbal ability tests typically fit in this category, especially when they are relatively short. However, a group-administered test will sometimes give the better estimate of the student's academic aptitude. This will happen if the group test samples more comprehensively from the target ability domain than the individually administered test. For example, the CogAT contains nine different fluid reasoning subtests: three verbal, three quantitative, and three figural/nonverbal. Because they try to measure a much broader sample of abilities, individually administered ability tests developed in recent years actually have far fewer reasoning subtests (Frazier & Youngstrom, 2007). In this case, the CogAT scores are generally more reliable (see Table 49.1), a better measure of g (Lohman, 2003a, 2003b), and an equally good or better measure success in school – even when method effects are controlled (e.g., CogAT predicting scores on a group-administered achievement test and the individually administered ability test predicting scores on an individually-administered achievement test).[7]

Errors of measurement in test scores. Of the many things that one might want to know about the scores on a test, surely one of the most important is the dependability of those scores. Measurement experts have long advocated that test users rely on the standard error of measurement (SEM) rather than the reliability coefficient when making inferences about the dependability

[7] Another common complaint about individually administered tests is that one has no way of knowing if the child did not understand the directions, or broke a pencil point, or responded inconsistently to different items ~~are~~ different subtests (e.g., Sattler & Hoge, 2006). This is true for every group test but Form 6 of CogAT. On this test, the pattern of each child's responses is compared to the expected pattern of responses given the total number of items at the child answered correctly. If the observed pattern of item or subtest scores is inconsistent with the expected pattern of item or subtest scores the confidence interval around the ability estimate for that child will be unusually wide. This indicates that there is considerable uncertainty about the score and that it should not be used to make high-stakes decisions about the child. If the amount of deviation from the expected pattern is extreme, a warning will be printed as well. In fact, these procedures were developed in response to the query of a parent about the failure of her son who took CogAT 5 to qualify for the G&T program.

Table 49.1 Standard errors of measurement for a 10-year old child scoring near the mean on several ability tests

	WISC-IV[a]	SB-V	OLSAT-8	Inview[c]	Raven[d] SPM	NNAT	CogAT 6
Verbal	3.9	3.6	5.7	5.3			3.4
Nonverbal/perceptual	4.2	3.9	5.8[b]	4.5[b]	3.0	6.1	3.7
Quantitative	4.5	5.3					3.3
Composite/full scale	2.8	2.8	5.7	3.5			2.2

Note: All SEM's on a scale with mean = 100, SD = 16

[a] Working Memory Composite used to estimate Quantitative for WISC IV.

[b] On OLSAT and Inview, the quantitative subtests are included in the nonverbal score. The proper comparison with CogAT is therefore with the CogAT QN partial composite. The SEM for the QN Composite is 2.7.

[c] Inview only reports SEMs for the individual subtests, not the three composite scores that are reported. SEM's for composite scores were estimated by $(\sum e^2/k^2)^{.5}$ (Feldt & Brennan, 1989). These were then converted to CSI scores (M 100, SD 16) using the norms tables.

[d] Estimated from Table RS3 147 and RS3 148 in Raven et al. (2000). Table RS3 147 shows approximate 67% confidence intervals for PR scores by age. These were then converted to a scale with M 100, SD 16 using Table RS3 148.

of test scores. The SEM makes it much easier to estimate the magnitude of score changes one is likely to see upon retest. To illustrate, Table 49.1 shows how errors of measurement vary across several commonly used ~~group~~ ability tests for children in grade 3. For ease of comparison, all scores are reported on a scale with mean 100 and standard deviation 16.

SEMs vary widely, both within a test (e.g., 3.6 for the Verbal Fluid Reasoning and 5.3 for the Quantitative factor index on the Stanford-Binet V) and between tests. In general, reliability increases with the number of items. For example, each level of the NNAT has 38 items. Students are allowed 30 min to attempt as many as they can. It has the largest SEM of any test in Table 49.1. On the Raven, however, students are given as long as they need (typically an hour or more) to attempt 60 items. Its SEM is less than half as large. But do the differences in SEM shown in the table matter? The 90% confidence interval for a student who receives a score of 100 on the NNAT is 88–112—a range of 24 points. For the CogAT Composite score it is 95.6–104.4—a range of 8.8 points.

This situation is actually substantially worse for scores at the extremes of the distribution, especially when scores approach the maximum possible score. This can happen on a group test when students answer most (or even all) of the items correctly. In such cases, errors of measurement will be smallest near the mean scale score and increase substantially at the extremes of the distribution. Commonly the SEM is from two to four times larger for very high scores than for scores near the mean. This can have disastrous consequences for efforts to identify gifted students—especially when scores are reported on an IQ-like scale rather than on the percentile-rank scale.[8] One way to reduce these errors of measurement is to test out of level. Put differently, one administers a level of the test that better matches the abilities of the student. This makes it less likely that the student will be unfairly penalized (or credited) for missing (or solving) a single item. Indeed the higher level of the test includes more difficult items which allows the student better opportunity to demonstrate his or her abilities. On CogAT, for example, users who need dependable scores for students in grades 3 and above who score above the 95th PR are advised to move up two test levels (Lohman & Hagen, 2002). This is easy to do because all tests at these grades have the same directions and time limits.[9]

[8] This is because percentile ranks are compressed at ~~details~~ of the distribution whereas scale scores spread out. For example, for tests such as CogAT and OLSAT that have a mean of 100 and standard deviation of 16, every SAS score above 134 receives the same percentile rank of 99.

[9] Out-of-level testing on tests such as CogAT and other tests that have vertically equated score scales has other benefits. For example, instead of administering the SAT or ACT to students with extremely high scores, one can more easily arrange to administer a higher level of the test. For example, level H on CogAT has a good ceiling even for able 12th graders. The universal scale score (USS) obtained by the student on the higher level of the test can easily be compared with multiple norm groups (e.g., for grades 6, 8, 10, and 12). This has the added advantage of keeping the student's scores on the same scale that was used for all of the other children who were administered the ability test. It is relatively easy to compute how many standard deviations such as score might be above the mean USS score for the student's age group. This would be a much better strategy then reverting to ratio IQs or making other highly questionable assumptions about the equivalence of the scale when scores are projected well beyond the range of data in the sample.

Individually administered tests are not immune, however. True scores (i.e., the score that would be obtained if the students could be tested many times without any memory of the previous test) are on average always closer to the mean than observed scores. The higher the score, the more likely it is to regress. Therefore, confidence intervals for extreme scores are always skewed toward the mean. The higher the score, the more substantial the skew (Stanley, 1971).

Measures of Motivation, Interest, and Creativity

Learning requires more than prior knowledge and good reasoning abilities. This is because learning is never a purely rational activity. Whether a child persists in thinking about something depends on affective and motivational factors. Sometimes affective engagement can be elicited by parents, teachers, and coaches. But more commonly high levels of engagement are better understood as a resonance or attunement between the child and the activity or domain. This fascination can be short lived or enduring. Either way, interest is a critical and easily measured aptitude for learning or performing well. Interest inventories can be helpful, especially for adolescents (see Lubinski, Benbow, & Ryan, 1995; Schmidt, Lubinski, & Benbow, 1998). For younger children, less formal methods may be used. For example, because children know more about topics that interest them, inventories of factual knowledge in particular domains can provide useful information about children's interests.

~~Finally, many~~ of the more important characteristics of students that function as aptitudes for learning are best obtained through ratings of trained observers. For example, motivation, creativity, and expressive communication skills can be estimated by well-trained teachers on rating scales such as the Scales for Rating the Behavioral Characteristics of Superior Students (Renzulli et al., 2002). Combining such information with information on interests, abilities, and achievements is still more art than science. Should teacher ratings be given greater weight than students' interests? And how should these measures be combined with estimates of ability and achievement?

Common Pitfalls

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

The Non-exchangeability of Measures

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

Suppose we administered two tests, plotted the scores for all students, and calculated the proportion that scored above a particular cut score (such as the top 5%). Next, we repeated the experiment using tests that showed different degrees of correlation. If we then plotted the proportion of students who were above various cut scores on both tests, we would get the curves shown in Fig. 49.1. The correlation between two tests is shown on the abscissa (x axis) of the figure. The correlation varies from $r = .5$ to $r = .975$. Plots for seven different cut scores are shown. These range from the selecting the top 1% (bottom curve) to the top 20% (top curve). For example, the bottom curve shows the proportion of students scoring in the top 1% of the distribution on one test who could also be in the top 1% of the distribution for the second test. The proportion ranges from only 13% when $r = .50$ to 76% when $r = .975$.

Figure 49.1 shows clearly that there will be considerable disagreement in classification among different tests unless the correlation is very high and the cut score is very low. Neither of these conditions typically applies. For example, the correlations among total scores (i.e., full-scale IQ scores) on individually administered ability tests range from $r = .68$ to $r = .85$. A common criterion is scoring in the top 3% of the distribution. Figure 49.1 shows that this means only about one-third to one-half of the students would be expected to score in the top 3% on both tests. Scores for shorter tests (e.g., verbal IQs) show lower correlations and would agree even less well. Achievement test scores show similar effects. For example, selecting students on the basis of their composite score on the ITBS misses approximately 40% of the students with the highest Reading Total scores. This is not what most people would expect for two variables that correlate $r = .91$.

Does this mean that any student who obtains a high score on a test should be considered gifted in the assessed domain? Not at all. A substantial part of the difference between the scores individuals obtain on two tests is due to the many influences we collectively call *errors of measurement*. If we were to administer a parallel form of the first test on a different occasion, then many students would not obtain high scores on both tests. The reliability coefficients of parallel forms rarely exceed $r = .90$. Even with this degree of reliability, Fig. 49.1 shows that only 60% of the students who are tested would score in the top 3% on both tests.

One implication, then, is that one should *never* base decisions about admission on a single score. Rather, one should average scores on parallel forms of a well-chosen test administered at different times. How much difference will an average score make?

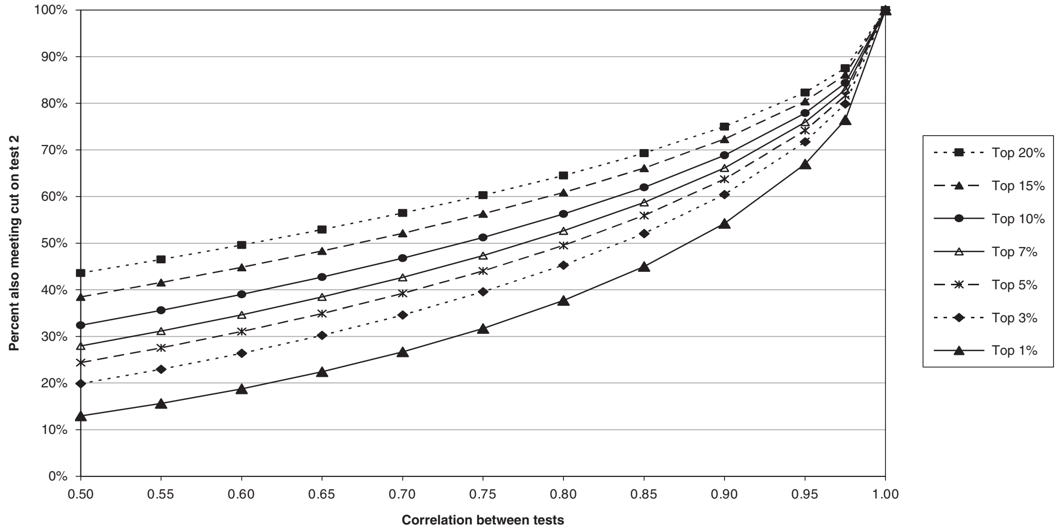

Fig. 49.1 Proportion of cases exceeding the same cut score on two tests, by the correlation between the tests

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

Small words can have large consequences. The shift between a rule that admits a student on the basis of a high score either on test 1 *or* on test 2 admits many more students than a rule that admits a student on the basis of a high score on both test 1 *and* test 2. The intermediate position—take the *average* of test 1 and test 2—admits yet a different group.

Figure 49.2 graphs the results of these three rules. Each panel shows a scatter plot of students' scores on the two tests. The left panel shows the students identified by the "and" rule, the center panel those identified by the "or" rule, and the right panel those identified by the "average" rule. Requiring students to score above a particular cut score on both tests 1 and 2 restricts the number of students who are identified. This is the effect of a two-stage screening process in which students must achieve a high score on the first test (e.g., a norm-referenced achievement test) and then a high score on a second test (e.g., an individually administered ability test). It is an unforgiving rule. A very high score on ~~one~~ test will not offset a score on the second test~~s~~ that is only slightly below the cut score.

Consider the case in which the cut score is set at the top 5% on both tests and the correlation between them is $r = .80$. Only about 50% of the students in the population who meet this criterion on one test will also meet it on the second test (see Fig. 49.1). This means that 50% of the 5% who met the criterion on test 1, or 2.5% of the total student population, will be admitted.

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

The disjunctive "or" rule is most defensible if the two tests measure different constructs such as language arts and mathematics. One should seek to identify students who excel in either domain, not just those who excel in both domains. If both tests measure the same construct, however, the statistically optimal rule is neither "or" nor "and" but rather "average." As Fig. 49.2 shows, the "average" rule will admit more students than the restrictive "and" rule and fewer than the liberal "or" rule. Essentially, students are admitted on the basis of where they fall on the 45° diagonal rather than on either the x axis or the y axis. Furthermore, using two scores to estimate ability (as in either the "and" or the "average" rule) substantially reduces regression effects. The only disadvantage of the "average" rule is that it requires averaging the scores! This is not as convenient as determining which students exceed a given percentile rank on each of the two measures.

Sometimes only a portion of the population is administered a screening test. In such cases, it is important to test a *much* broader segment of that population than just those students who score highly on an achievement test or who are nominated by their teachers. Testing only those who meet the desired criterion on an achievement test (e.g., 97th PR) will eliminate the majority students who score higher on the ability test than on the achievement test. Furthermore, testing only those students nominated by teachers does not reduce the amount of regression to the mean that will be observed when students are tested the second time (Lohman & Korb, 2006).

Testing more children is much easier to do when using a group-administered test rather than an individually administered test. A reasonable rule is to test every child who scores above average on one or more

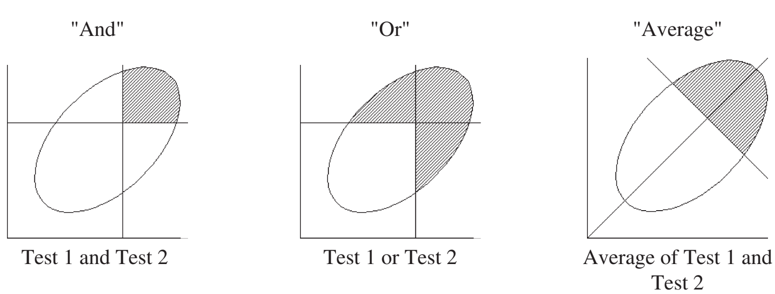

Fig. 49.2 Plots of the effects of three selection rules: (**a**) high scores on test 1 *and* test 2; (**b**) high scores on test 1 *or* test 2; and (**c**) high scores on the *average* of test 1 and test 2

of the major batteries of the achievement test (typically Reading Total and Math Total for elementary students). If this is done, it is often more administratively convenient to administer the ability test to all children. Classroom teachers are most likely to go along with this procedure if the ability test does more than identify the most able students but also provides information that they can use to help all students learn.[10]

Combining Scores from Different Tests

Combining scores from different tests is thus almost always a better policy than using a single score. But how should scores be combined? Here are a few guidelines. Specific procedures are provided later in this chapter.

1. *If scores come from the same test* (e.g., ITBS Reading Total scores for grades 3 and 4), *average the scaled scores for the two administrations of the test.* For example, on the ITBS these are called Standard Scores. For CogAT they are called Universal Scale Scores. On CogAT and other ability tests, one can also average the Standard Age Scores (SAS).
2. *Expect that averaged scores will regress to the mean.* Unless norms for partial composites are provided, one cannot use norms tables to look up the percentile ranks of the averaged scaled scores in the same way that one looks up the percentile ranks of individual test scores. Similarly, averaged SAS scores will not have the same PR associated with each score as individual SAS scores. Therefore, base decisions on rank within the local group of all students' average scale scores, not on whether these average scores exceed a fixed percentile rank that applies only to individual scores.
3. *If scores come from tests that use different score scales, first put them on a score scale that has the same mean and standard deviation*. Use the Normal Curve Equivalent (NCE) scores that are provided for some tests. If NCE's are not available, then one can convert scaled scores to *z*-scores by computing the mean and standard deviation for each set of scaled scores. These *z*-scores have a mean of 0 and a standard deviation of 1. The standard deviation is one measure of the spread or dispersion of scores. By making the standard deviations the same, we insure that one variable does not overwhelm the other when they are combined. For example, to combine ITBS Reading Total Standard Score (SS) and CogAT Verbal Battery Universal Scale Score (USS), first get the mean and standard deviation for each set of scores. Then compute *z*(reading) and *z*(Verbal Battery) using a function such as "standardize" in Microsoft Excel. Finally, add the two *z*-scores together to get a composite score that weights each of the component scores equally.
4. *Generally avoid combining percentile ranks (PRs)*. Use scale scores instead. PR scores make only crude distinctions at the top of the score scale. For example, when the mean is 100 and SD 16, every score above 134 is at the 99th percentile. Furthermore, the average of two PR's is generally not the same as the PR of the average of the two scale scores.
5. *If possible, base the weights assigned to different tests on research*. Although equal weights work well when combining ability and achievement test scores, one would not want to give equal weight to scores that either are less reliable or have weaker relationships with outcome measures. For example, even though interest and persistence are critical, measures of these constructs are much less reliable and show at best moderate correlations with outcomes. It would not be appropriate to weight them the same as measures of achievement or ability.

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

[10] CogAT (Form 6) provides this sort of information. Score reports contain a profile index that summarizes the level and pattern of scores across the three batteries for every student. This profile is linked to specific suggestions on how to adapt instruction (see www.cogat.com).

Identifying Academically Talented Minority Students

Prediction of Achievement for Minority Students

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

Several researches have investigated this issue. All have found that the abilities that best predict achievement in reading, mathematics, social studies, and science are the same for White, Black, Hispanic, and Asian-American students (Keith, 1999; Lohman, 2005b). For example, verbal reasoning abilities are the strongest predictor of reading comprehension for all ethnic groups. Nonverbal reasoning contributes least to the prediction. Indeed, nonverbal reasoning abilities sometimes have a significant *negative* regression weight in the prediction of achievement once verbal and quantitative reasoning abilities are in the equation (Case, 1977; Lohman, 2005b). This means that some students with high nonverbal reasoning scores are actually *less* likely to achieve in school than are other students with similar levels of verbal and quantitative abilities.

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

Predictions about future performance assume that a student's rank within the comparison group on the aptitude test will remain relatively constant over time. This does not mean one assumes that scores are fixed. Scores that report rank within age or grade group easily mask the fact that all abilities are developed; all respond to practice and instruction. Rather, the assumption is that a student's rate of growth on the skills measured by the test will be the same as other students in the norm group who obtained the same initial score. This is unlikely either if the student's experiences to date differ from those of the norm group or if her subsequent experiences depart from the norm. For example, lack of experience in a domain will lead to a lower initial rank than will be achieved later as the student has the necessary learning experiences. This is especially true for well-defined skill sets that are quickly learned (e.g., learning the letters of the alphabet) rather than for open-ended skill sets that require extensive practice (e.g., verbal comprehension).

A student can also fall behind over time by improving, but at a slower rate than her peers. This is one reason why some aptitude tests over-predict the future academic performance of some minority students (Willingham, Lewis, Morgan, & Ramsit, 1990). Therefore, programs that aim to assist minority students in developing their academic talents might best understand their task as one of falsifying a prediction about growth rate. This is not easily done. Contrary to popular myth, complex skills and deep conceptual knowledge do not suddenly emerge when the conditions that prevented or limited their growth are removed (cf. Humphreys, 1973). The attainment of academic excellence comes only after much practice and training. It requires the same level of commitment on the part of students, their families, and their schools as does the development of high levels of competence in athletics, music, or other domains of nontrivial complexity. Furthermore, because the relationship between aptitude and outcome is probabilistic, one cannot expect that every student who is identified as likely to succeed will do so. The critical issue for programs that aim to assist these children, however, is to maximize the proportion of identified students that do succeed.

Judging Test Bias by Mean Differences Rather Than by Predictive Validity

A selection policy that uses either ability or achievement tests alone or that combines, say, mathematics achievement and quantitative reasoning ability would select proportionately fewer Black and Hispanic students than White and Asian-American students. How, then, can one attend to the relevant aptitude variables and increase the representation of minority students served? All recognize that many students—especially those whose first language is not English—have not had the same opportunities to develop skills in the English language. Therefore, many schools screen students with nonverbal tests, teacher questionnaires, and performance assessments because differences between ELL and native speakers of English are sometimes smaller on such tests.[11]

The need for a test that minimizes group differences is a consequence of the assumption that one must always compare every student to every other student in an age or grade cohort. But this is no more necessary (or desirable) than comparing all primary children with each other rather than with others in the same grade. There are many reasons for the assumption that all students must be compared to the same norm group. In part, it stems from the laudable desire to be fair. All children are compared to the same standards, or so it seems. In part, it stems from the failure to appreciate the extent to which the norm group to which a child's score is compared often changes monthly on ability tests and weekly on achievement tests to accommodate differences in children's experiences in the culture or in school. And in part, it stems from the administrative convenience of using norms provided by the publisher rather than having to develop local or local subgroup norms. Other things being equal, we surely would prefer the assessment procedure that showed the smallest difference between ethnic groups. However, other things are rarely equal.

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

1. *Reinforces the tendency to interpret intelligence and other ability tests as measuring innate abilities.* If scores on ability tests (including nonverbal tests) depend on background and education, then one must take these factors into account when interpreting them. The alternative—to interpret test scores as measures of innate abilities largely unaffected by such factors—avoids these complications. Thus, the decision to use a common cut score on aptitude tests inadvertently encourages the naïve but false belief that ability tests measure innate rather than developed abilities.
2. *Encourages the use of less reliable tests.* Other things being equal, group differences will be smaller on less reliable tests than on more reliable tests. For example, performance tests are generally less reliable than objective tests and will generally show smaller group differences than objective tests that measure the same abilities. In the extreme, a completely unreliable test will show no differences between groups even when true differences are large. Therefore, evaluating tests by the extent to which they achieve the goal of proportional representation will tend to favor shorter and otherwise less reliable tests over longer and more reliable tests.
3. *Encourages the use of old tests with outdated norms.* More students (minority and majority) will attain high scores on a test with outdated norms than on a test with recent norms. This was demonstrated above for the out-of-date U.S. norms commonly used for interpreting scores on the Progressive Matrices Test.
4. *Encourages the lowering of standards.* If the scores for two groups differ in their means and show equal variability, then lowering the admission standard will increase the proportion of students from the lower-scoring group. For example, assume one changes the cut from, say, the top 5% to the top 10% of cases. There will be a greater proportion of students from the lower-scoring group when the cut is set at the top 10% rather than at the top 5%. Note, however, that the total number of students admitted has now doubled.
5. *Encourages the use of less valid tests.* The hope that one can use a common cut score for all applicants

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

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

The Need for Within-Group Comparisons

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

This does not mean that every child needs to be compared only to the handful of other children in the population who share her unique circumstances. Simply comparing an ELL student to all other ELL students in a grade will help enormously. If the sample is sufficiently large, then one can further subdivide into groups with little, intermediate, and extensive exposure to the English language (Ortiz & Ochoa, 2005). One of the most important advantages of group-administered ability tests is that they allow these sorts of comparisons when a common test is administered to all of the students in a school or district. Furthermore, one need not derive formal norms to make such within-group comparisons. A simple rank order of scores will often serve the purpose (see http://faculty.education.uiowa.edu/dlohman/doc/Sample_data_set3.xls for examples).

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

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

Two Identification Procedures

I have argued that the best way to identify students who are likely to excel in particular domains is to measure the specific aptitudes that are most needed for successful learning in those domains under the instructional systems that are available or can be developed. I have also argued for greater diversity in the programming options than historically have characterized programs for the gifted. In this section, I show how identification systems of this sort can be developed. See Lohman and Renzulli (2007) for an approximation to these procedures that first converts percentile ranks or IQ-like scores to single digit numbers and then combines them.

How to combine different kinds of information is a critical issue when identifying gifted children. Arraying this information in a matrix makes it simultaneously available, but does not offer a principled way to combine it. Some programs prefer to follow traditional identification practices in which children are identified primarily (or solely) on the basis of ability and/or achievement test scores that are unusually high, using either national or local norm groups. Others have argued that programs should also serve children whose test scores are somewhat lower (e.g., the top 20% in the local group) but whom teachers believe exhibit unusual creativity, commitment to learn, or accomplishments in particular domains (Renzulli, 2005). The identification procedures described below allow both of these perspectives.

The first procedure shows how to determine a child's standing on any score (or combination of scores) in three norm groups: the nation, the local population, and opportunity-to-learn subgroups within that population. The procedure works best when all students in a particular grade in the local population (i.e., school or district) are administered the screening test. It also requires that one know the basics of using a spreadsheet application (such as Microsoft Excel).

The second procedure shows how to combine ability, achievement, and teacher ratings in a principled way. The first step is to combine scores on ability and achievement tests into two composite scores: a verbal-reading composite and a quantitative-figural-mathematics composite. Then these composite scores are compared with teacher ratings to inform decisions about acceleration or enrichment for those who have high scores on one or both composites.

These procedures are illustrated using the CogAT—Form 6 (Lohman & Hagen, 2001), the Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001) and the Scales for Rating the Behavioral Characteristics of Superior Students (SRBCSS; Renzulli et al., 2002). Other combinations of tests and rating scales may be used. Although helpful for other purposes, there is no need for the ability and achievement tests to be co-normed. For rating scales and other observational techniques, it is critical that raters be trained. Ratings provided by untrained raters commonly suffer from halo effects (i.e., giving consistently high or low ratings to a student across different items or dimensions). One indication of this is given by the correlations among ratings for dimensions such as learning ability, creativity, and motivation. For example, correlations between creativity and academic ability or achievement are modest when actual behaviors are observed. If ratings of students on these dimensions show high correlations, then one should suspect halo effects. Halo effects should also be suspected if measures of internal consistency (e.g., coefficient alpha) for individual scales are extremely high (i.e., greater than .80). For example, it is well known that creativity in art, music, and academics shows only weak correlations. If the creativity scale asks for ratings of students in these different domains, then a high alpha indicates that raters gave particular students similar ratings across all domains. This invalidates the ratings.

Decisions about educational programming—especially whole grade acceleration—would require information in addition to the achievement test scores, ability test scores, and teacher ratings discussed here. These would typically include some measure of the student's interests, social skills, anxiety, and other characteristics that would be expected to influence the probability of successful learning in the different educational placement options under consideration (Assouline, Colangelo, Lupkowski-Shoplik, Lipscomb, & Forstadt, 2003).

Procedure 1. Multiple Norm Groups, Multiple Perspectives

The basic idea here is that it is useful to examine the performance of a child from multiple perspec-

Book_Shavinina_9781402061615_Proof2_December 2, 2008

tives. Here I emphasize issues that are salient when attempting to identify the most talented poor and minority students. However, the scores of any child can be compared to those of many different groups (e.g., current grade and a prospective higher grade). Detailed directions for using this method are provided in Lohman (2006) and in the sample data set that accompanies that monograph. Here are the steps for using only one test score. Examples in the sample data set that is available on the website show how to combine scores on two test scores (such as mathematics achievement and quantitative reasoning abilities). This is important because identification procedures that average ability and achievement scores for particular domains better identify not only those who currently excel but also those who are most likely to continue to excel (see Lohman & Korb, 2006).

Step 1—Preparing the data. Get the required data into a spreadsheet. For each student, this would include the student's name or ID, an opportunity to learn index (such as ELL status), national percentile ranks (PRs), or other norm-referenced test scores. If different test scores will be averaged, then scale scores should be recorded as well. For ability tests such as CogAT, these Standard Age Scores (SAS) for one or more of the test batteries could be used instead of scale scores.

Step 2—Getting local ranks. Sort the data by percentile ranks (or SAS scores). This will provide local ranks.[12] Local score distributions generally provide a better way to determine which students are most likely to be mismatched with the instruction they are receiving than will national norms. They also make it much easier to identify a relatively consistent number of students across years.

Step 3—Looking within groups defined by opportunity to learn. Sort the data again by opportunity to learn (as the first sorting variable in Excel) and then PR or SAS (as the second sorting variable in Excel). For example, if two opportunity-to-learn groups are used (e.g., ELL versus native speakers), then the most talented ELL students will be those with the highest ranks within the first group and the most talented native-speaking students will be those with the highest ranks in the second group. What kind of enrichment or acceleration to suggest for each depends on the students' levels of achievement and on other factors (such as interest, motivation, and the availability of different educational programs).

[12] Note that ranks are not the same as the percentile ranks provided in norm tables. However, for most purposes, a simple rank order of the scores is all that is needed. Other scores (e.g. Standard Age Scores) provide additional information on the size of the score gaps between students with different ranks.

Procedure 2. Using Ability Test Scores and Teacher Ratings

This example uses all three CogAT batteries, reading and mathematics achievement test scores, and the three main scales from the Scales for Rating the Behavioral Characteristics of Superior Students (*SRBCSS*; Renzulli et al., 2002). The three scales from the SRBCSS are Learning Ability, Motivation, and Creativity.

How best to combine scores from the three *CogAT* batteries with measures of current achievement when predicting future academic success is well documented. Competence in a broad range of verbal domains (e.g., literary arts, history) is best predicted by the CogAT Verbal SAS score and the ITBS Reading Total scale score. As noted above, either use NCE scores or convert scale scores to *z*-scores by computing the mean and standard deviation for each set of scaled scores. These *z*-scores have a mean of 0 and a standard deviation of 1. By making the standard deviations the same, we insure that one variable does not overwhelm the other when they are combined. Then simply average these two scores.

On the other hand, success in mathematics and domains of study that demand quantitative thinking is best predicted by a combination of the CogAT Quantitative and Nonverbal Reasoning Batteries and the ITBS Mathematics Total scale score. The CogAT Quantitative-Nonverbal (QN) Composite captures the ability dimension. As before, use NCE scores or convert the QN Composite SAS score to a z-score and then average it and the z-score for the ITBS Mathematics Total scale score. This gives two broad ability-achievement composites: verbal-reading and quantitative-figural-mathematics. Students should be identified if they have high scores on either dimension.

Research with the *SRBCSS* shows that each of its three main scales provides unique information.

Fig. 49.3 Scheme for combining ability, achievement, and teacher ratings to create a talent pool

		Teacher Rating of Learning ability, Motivation, or Creativity	
		Below Average	Above Average
Verbal-Reading Composite or Quant-Nonverbal-Math Composite	>97th PR	II Admit but watch	I Admit
	>80th PR	IV Test next year	III Enrichment

Therefore, teacher ratings should be considered high if any one of the three ratings is high.[13]

The identification scheme is shown in Fig. 49.3. The vertical dimension distinguishes children who exhibit superior abilities from those who exhibit above-average abilities. I have arbitrarily set the cut scores as scoring at or above the 97th local percentile rank or at or above the 80th local percentile rank on either the verbal-reading or the quantitative-nonverbal-mathematics composite. The first criteria is commonly used in gifted programs; the second is recommended when casting a talent broader net (Renzulli, 2005). The horizontal dimension distinguishes between children who, when compared to other children *nominated* for the program, obtain above-average teacher ratings on any one of the three SRBCSS scales and students who obtain average or below-average teacher ratings.[14] Note that, for ratings, the average is computed only on the subset of the student population who are nominated for inclusion in the program. Combining these two criteria gives four categories of assessment results.

Children in category I exhibit superior reasoning and achievement and are rated as highly capable, motivated, or creative by their teachers. Children in category II also exhibit superior reasoning and achievement but, when compared to other children who were nominated, are not rated as highly by their teachers on any one of the three major scales of the SRBCSS. Programs that follow a traditional identification scheme (e.g., self-contained classrooms or schools) would accept children in category I. Most would also accept children in category II, especially if it is difficult to defend rejection on the basis of low teacher ratings. However, if this is done, then the progress of children in category II should be monitored more closely. Children in category III exhibit somewhat lower but strong reasoning abilities (80th–96th PR) on one of the ability-achievement composites and are rated as highly capable, motivated, or creative by their teachers. These children would be included in school-wide enrichment programs that aim to serve a broader range of children. Schools with highly diverse student populations would find that many of their best poor and low-achieving students would fall in this category. Combining test scores and ratings in this way would enable these schools to identify the students most likely to benefit from curriculum compacting or enrichment programs, including instruction at a higher level than that received by most other students in the school. Finally, children in category IV exhibit good but not exceptional abilities (between 80th and 96th PR) and are not rated as unusually capable, motivated, or creative by their teachers. Although good students, these children would not be provided with special programming on the basis

[13] Because ratings are strongly correlated, accepting all students with an "above average" rating on any one of the three rating scales will identify considerably more than half of the students. This is undesirable only if schools cannot find ways to provide enrichment or other instruction for these students. Furthermore, if possible, ratings on every student should be obtained from more than one teacher and then averaged.

[14] Because one is using an "or" rule, considerably more than half of the students will be in categories I or III than in II or IV. If the correlations among the scales are known, one could estimate this proportion from Fig. 49.1. If a smaller fraction of the population is desired, then the selection rule could be changed to, say, top quartile versus all others on any one of the three rating scales.

of either their ability-achievement scores or teacher ratings.[15]

Suggestions for Policy

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

Although adapting instruction to the needs of students is a critical aspect of any successful program, I have not emphasized it here. Too often, children are labeled "gifted" on the basis of an IQ test; other affective and cognitive aptitudes required for success are ignored or are only considered after the student has been identified. But giftedness means superior aptitude or talent for something, not for everything. Programs would do a better job of identifying talented children if they started with a clear understanding of the types of expertise that they are able to develop and the demands of the educational programs that they can offer to develop talent in these domains. Together, these will more clearly define the personal characteristics that will function as aptitudes for success in those programs.

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

How can educators implement a policy consistent with the principles outlined here? Here are some final suggestions:

1. *Define the purpose(s) of the TAG program.* Is the emphasis on *T* (talent) or *G* (gifted)? Is the goal to identify and serve those students who demonstrate unusually high levels of academic ability and accomplishment? If so, then traditional procedures for identifying and serving academically gifted students can be used. Poor and minority students will be included in this group, although not at a level that approaches their representation in the population. Attempts to achieve greater minority representation by using nonverbal tests and other measures that are not good measures of scholastic aptitude

---

[15] One could combine this procedure with the multiple norm group procedure, especially for identifying ELL students who show promise in the verbal-reading domain.

will indeed include more ELL students in the program. Unfortunately, these will often not be the most academically promising students. On the other hand, if the goal is to identify the most academically talented students in underrepresented populations regardless of current levels of academic attainment, then procedures like those outlined in this chapter will be more successful. However, options for educational placement and programming will need to be much more diverse than is currently the case. Perhaps in this way, TAG programs could infuse procedures for identifying academic talent and then providing developmentally appropriate instruction into mainstream educational practices. It is not only academically gifted students who are not well served by a rigidly age-tracked educational system.

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

3. *Obtain the most reliable and valid measures of achievement, reasoning abilities, and other aptitude variables for all students.* Whenever possible, measure the behavior of interest rather than something that merely predicts that behavior. If interested in children's weight, then weigh them. Do not measure their heights and try to predict weight from height. Similarly, if the goal is to identify students who have unusual talent for particular academic domains, obtain measures of domain-specific achievement, the student's ability to reason in the symbol systems required for new learning in that field of study, interest in the domain, and persistence under similar instructional conditions. For example, to identify students who excel in mathematics, first measure mathematics achievement using a well-constructed, norm-referenced achievement test that emphasizes problem solving and concepts. To identify students who are most likely to show the strongest future development, combine scores on the mathematics achievement test with scores on measures of quantitative and figural reasoning abilities. Combine the scores in a way that weighs mathematics achievement and reasoning abilities equally. To assess interests, inquire specifically about the students' interests in mathematics or in occupations that require mathematical thinking. Interest inventories can be helpful, especially for adolescents (see Lubinski et al., 1995). Finally, estimate persistence, anxiety, and other important affective aptitudes from ratings obtained from teachers and others who have worked with the child in situations similar to those in the planned acceleration program. Keep in mind that aptitude can only be estimated when a student's performance on a task is compared with the performance of other students who have had roughly similar learning opportunities. Common cut scores on less valid and reliable tests may identify significant numbers of minority students, but many of them are not the students who have the greatest academic talent.

4. *When identifying students, make better use of local norms on both ability and achievement tests, especially when identifying students whose accomplishments in particular academic domains are well above those of their classmates.* On

norm-referenced tests, examine local percentile ranks for particular domains such as mathematics or science in addition to national percentile ranks for these scores. In general, do not make decisions on the basis of composite scores. When making instructional placements, use local norms to determine the appropriateness of the match. For example, if a student will be placed with seventh graders for mathematics, compare her performance on a test with seventh grade mathematics content to the performance of students in the prospective seventh-grade class.

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

6. *Remind policy makers that professional judgment is required.* There is no foolproof way to identify those children who will develop the highest levels of academic excellence in adolescence or the highest levels of professional expertise as adults. Simple schemes that establish an arbitrary cut score on an IQ or achievement test are administratively convenient but identify only a fraction of those who will later attain excellence. Furthermore, such schemes necessarily disadvantage children who have had fewer opportunities to develop the abilities measured by the tests at the time selections are made. Until identification policies better match our understanding of how academic talent is expressed and developed, the field will struggle to get beyond present practices, many of which can identify only that small fraction of the population that currently excel on all dimensions that are assessed.

References

Ackerman, P. L. (2003). Aptitude complexes and trait complexes. *Educational Psychologist, 38*, 85–94.

Assouline, S., Colangelo, N., Lupkowski-Shoplik, A., Lipscomb, & Forstadt, L. (2003). Iowa Acceleration Scale (2nd edition). Scottsdale, AZ: Great Potential Press.

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York: Harper & Brother.

Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Mind, brain, experience, and school*. Washington, DC: National Academy Press.

Carroll, J. B. (1974). The aptitude-achievement distinction: The case of foreign language aptitude and proficiency. In D. R. Green (Ed.), *The aptitude-achievement distinction* (pp. 286–303). Monterey, CA: CTB/McGraw-Hill.

Case, M. E. (1977). *A validation study of the Nonverbal Battery of the Cognitive Abilities Test at grades 3, 4, and 6.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

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

Colangelo, N., Assouline, S. G., & Gross, M. U. M. (2004). *A nation deceived: How schools hold back America's brightest students* (Vol. 1 & 2). Iowa City, IA: The University of Iowa, The Connie Belin & Jacqueline N. Black International Center for Gifted Education and Talent Development.

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

Cronbach, L. J. (1976). Measured mental abilities: Lingering questions and loose ends. In B. D. Davis & P. Flaherty (Eds.), *Human diversity: Its causes and social significance* (pp. 207–222). Cambridge, MA: Ballinger.

Ericsson, K. A., Nandagopal, K., & Roring, R. W. (2005). Giftedness viewed from the expert performance perspective. *Journal for the Education of the Gifted, 28*, 287–311.

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

Frazier, T. W. & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence*.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Book, *35*, 169–182.

Gardner, H. (2003, April). *Multiple intelligences after twenty years*. Paper presented at the meeting of the American Educational Research Association, Chicago. Retrieved August 11, 2005, from Harvard University, Project Zero Website: http://www.pz.harvard.edu/PIs/HG.htm

Glaser, R. (1992). Expert knowledge and processes of thinking. In D. F. Halpern (Ed.), *Enhancing thinking skills in the sciences and mathematics* (pp. 63–75). Hillsdale, NJ: Lawrence Erlbaum.

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

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). *Iowa Test of Basic Skills: Form A*. Itasca, IL: Riverside.

Horn, J. L., & Blankson, N. (2005). Foundations for better understanding cognitive abilities. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 41–68). New York: Guilford Press.

Humphreys, L. G. (1973). Implications of group differences for test interpretation. In *Proceedings of the 1972 Invitational conference on testing problems: Assessment in a pluralistic society* (pp. 56–71). Princeton, NJ: ETS.

Humphreys, L. G., & Davey, T. C. (1988). Continuity in intellectual growth from 12 months to 9 years. *Intelligence, 12*, 183–197.

Hunt, E. (2000). Let's hear it for crystallized intelligence. *Learning and Individual Differences, 12*, 123–130.

Jackson, N. E. (1988). Precocious reading ability: What Does It Mean? *Gifted Child Quarterly, 32*, 200–204.

Jackson, N. E. (2003). Young gifted children. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 470–482). Boston: Pearson Education Inc.

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White test score gap*. Washington, DC: Brookings Institution.

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

Lohman, D. F. (2003a). *The Wechsler Intelligence Scale for Children III and the Cognitive Abilities Test (Form 6): Are the general factors the same?* http://faculty.education.uiowa.edu/dlohman/pdf/CogAT-WISC_final_2col2r.pdf

Lohman, D. F. (2003b). *The Woodcock-Johnson III and the Cognitive Abilities Test (Form 6): A concurrent validity study*. http://faculty.education.uiowa.edu/dlohman/pdf/CogAT_WJIII_final_2col%202r.pdf

Lohman, D. F. (2005a). *Identifying academically talented minority students*. (RM05216). Storrs, CT: National Research Center on the Gifted and Talented, University of Connecticut.

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

Lohman, D. F. (2006b). Beliefs about differences between ability and accomplishment: From folk theories to cognitive science. *Roeper Review*, *29*, 32–40.

Lohman, D. F., & Hagen, E. P. (2001). *Cognitive Abilities Test (Form 6).* Itasca, IL: Riverside.

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

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

Lohman, D. F., Korb, K., & Lakin, J. ~~(in press).~~ Identifying academically gifted English language learners using nonverbal tests: A comparison of the Raven, NNAT, and CogAT. ~~Gifted Child Quarterly.~~

Lohman, D. F., & Renzulli, J. S. (2007). *A simple procedure for combining ability test scores, achievement test scores, and teacher ratings to identify academically talented children.* http://faculty.education.uiowa.edu/dlohman/pdf/Draft%20Lohman-Renzulli%20ID%20system%20for%

Lubinski, D., & Benbow, C. P. (2000). States of excellence. *American Psychologist, 55*, 137–150.

Lubinski, D., Benbow, C. P., & Ryan, J. (1995). The stability of vocational interests among the intellectually gifted from adolescence to adulthood: A 15-year longitudinal study. *Journal of Applied Psychology, 80*, 196–200.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 136–181). New York: Guilford Press.

Naglieri, J. A. (1997). *Naglieri nonverbal ability test: Multilevel technical manual.* San Antonio, TX: Harcourt Brace.

Nickerson, R. S. (1998). Confirmation bias: Ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 410–433.

Ortiz, S. O., & Ochoa, S. H. (2005). Advances in cognitive assessment of culturally and linguistically diverse individuals. In D. P. Flanagan & P. L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, Tests, and Issues* (2nd ed., pp. 234–250). New York: Guilford Press.

Presseley, M., Raphael, L., Gallagher, J. D., & DiBella, J. (2004). Providence-St. Mel School: How a school that works for African American students works. *Journal of Educational Psychology, 96*, 216–235.

Raven, J. C. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1–48.

Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's progressive matrices and vocabulary scales, section 4: Advanced progressive matrices, sets I and II.* London: H. K. Lewis.

Raven et al. (2000). Manual for Raven Progressive Matrices and Vocabulary Scales: Research Supplement No. 3. San Antonio: Harcourt Assessment.

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

Renzulli, J. S., Smith, L. H., White, A. J., Callahan, C. M., Hartman, R. K., & Westberg, K. L. (2002). *Scales for rating* *the behavioral characteristics of superior students*. Mansfield Center, CT: Creative Learning Press.

Roid, G. (2003). *Stanford Binet Intelligence Scales, technical manual* (5th ed.). Itasca, IL: Riverside.

Ruf, D. L. (2003). *Use of the SB5 in the assessment of high abilities* (Stanford-Binet Intelligence Scales, Fifth Edition, Assessment Service Bulletin No. 3). Itasca, IL: Riverside.

Sattler, J .M., & Hoge, R. D. (2006). *Assessment of children: Behavioral, social, & clinical foundations* (5th ed.). La Mesa, CA: Jerome M. Sattler.

Schmidt, D. B., Lubinski, D., & Benbow, C. P. (1998). Validity of assessing educational-vocational preference dimensions among intellectually talented 13-year-olds. *Journal of Counseling Psychology, 45*, 436–453.

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

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

Stanley, J. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Tannenbaum, A. J. (1965). [Review of the IPAT Culture Fair Intelligence Test.] In O. K. Buros (Ed.), The Sixth Mental Measurements Yearbook (721–723), Highland Park, NJ: Gryphon Press.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science, 14*, 623–628.

VanTassel-Baska, J. (2000). The on-going dilemma of effective identification practices in gifted education. *The Communicator, 31*, 39–41.

Willingham, W. W., Lewis, C., Morgan, R., & Ramsit, L. (1990). *Predicting college grades: An analysis of institutional trends over two decades*. New York: The College Board.