

**CHAPTER 12**  
**ABILITY TESTING AND TALENT IDENTIFICATION**

**David F. Lohman**  
**Megan Foley Nicpon**  
**The University of Iowa**

**Draft of a chapter to appear in:**

In S. L. Hunsaker (Ed.), *Identification of students for gifted and talented services: Theory into practice.*

## BACKGROUND AND THEORETICAL FOUNDATION: DEFINING GIFTEDNESS

Identifying something presupposes understanding what that something might be. Giftedness is no exception. Unfortunately, the term *gifted* has as many definitions as there are theories of giftedness (Kaufman & Sternberg, 2007). For some, giftedness means high general ability (*g*). For others, it means the promise of excellence in a domain in which achievements are valued by a society. For yet others, real giftedness means making extraordinary contributions to a field as an adult. Some include creativity in the definition, either as an aspect of giftedness that all gifted students would be expected to display, or as a unique form of giftedness. Others include constructs such as leadership, practical intelligence, music, art, and athleticism.

When people define a word differently, their conversations are prone to miscommunication and conflict. Even if the staff of a talent identification and development program share a well-articulated definition of the term *gifted*, other teachers in the school, the school administrators, and parents will surely have different understandings. Often, these beliefs will be inconsistent with much that experts know about giftedness (Lohman, 2006a). Most non-experts would define giftedness as innate, general cognitive ability. This is no longer the view of experts in the field. For this reason, some of these experts suggest replacing the word “*gifted*” with a less value-laden and misunderstood term (Borland, 2004; Renzulli, 2005). For example, instead of a *gifted* program, one might speak of a talent identification and development program.

Too often, students are first labeled “*gifted*” and only then do program administrators worry about what to do with them. A talent identification and development perspective avoids this problem because it more clearly binds the identification process to educational programming. If a child is said to have mathematical talent or musical talent, or both, then the only question is how best to develop those talents. A creative writing program would not be the most appropriate match. Talents of greatest interest to school personnel are those needed to develop the forms of expertise sanctioned by society through the formal training systems made available in its schools.

A talent perspective also makes it easier to identify the particular interest, motivation, perseverance, and other personal characteristics that will be needed for talent development. Efforts to assist a child in developing her musical abilities are unlikely to succeed if the child is uninterested in learning music or unable to persist in the study of it. By focusing on specific talents, one also makes it more likely that schools will value a broader range of outcome measures that can index talent development, such as above-level test scores, individual performances, and regional competitions. In this way school personnel are more likely to recognize behaviors and accomplishments that indicate unusual creativity in the specific talent domain.

## TALENT DEVELOPMENT AND EXPERTISE

Talent development can be understood as the process of acquiring expertise in a domain. At some point, the most talented and dedicated individuals attain levels of competence that are exhibited by only a handful of other people. But getting there requires many years of sustained effort on the part of the learner and a parallel commitment by teachers and coaches. Understanding the magnitude of the task helps clarify the importance of engaging the child in learning activities that directly assist in developing that expertise. For example, helping students develop mathematical talent or musical talent means providing advanced instruction in mathematics or music rather than exercises in thinking skills. It means choosing projects and other enrichment activities that are not only interesting and enjoyable, but also contribute to the attainment of competence in the domain. This is especially important for school-based programs that serve students whose parents cannot afford summer classes or other focused talent-development activities. These students must rely almost exclusively on the school to assist in the development of their talents.

Identifying academically talented children thus depends not only on a clear definition of giftedness, but also on the range of educational programs that will be made available to children who are selected to participate in them. Programs range from highly selective schools for profoundly gifted children to enrichment activities for all children in a school who show interest in a particular activity and the ability to engage meaningfully and productively in it. Clearly, the 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 Schoolwide Enrichment Model (Renzulli, 2005). Nevertheless, the same principles govern each. It is these principles that are the focus of this chapter.

## TALENTS SCHOOLS SHOULD DEVELOP

At the most general level, identifying talent means discovering or recognizing an individual's predisposition to acquire competence in some domain. Deep learning in that domain proceeds at a more rapid pace for the talented individual than in otherwise similar individuals. Potentially, then, there are as many different talents as domains of human competence. Some domains are broader than others. For example, mathematics is a much broader domain than calculation. To show remarkable abilities in mental calculation does not mean that one has great potential for mathematics. Similarly, some domains are more highly valued by society: remarkable penmanship is less valued in schools today than creative writing skills; this was not always the case. Some talents have instrumental value to society whereas others are deemed to be of inherent value and are often pursued as an avocation. For example, surgical skills have instrumental value; people usually do not learn to be surgeons just for their personal enrichment or edification. Musical skills however, are often developed for their own sake.

Given the diversity of talents, it is useful to distinguish among talents that schools are (or should be) prepared to develop, talents that schools can encourage through their connections with external organizations, and talents that are outside of the schools' purview. Each category is described in Box 1.

A broad focus on talent identification and development both liberates and complicates. We move from a single clearly defined and well-measured talent (general academic aptitude) to a much larger set of talents, some of which are poorly defined and difficult to measure. Although the diversity of talents that schools recognize and develop should, by no means, be limited by current practices, at the very least, schools should be able to recognize and develop talents in those domains in which they already have programs, some of which are extensive and well-funded. Van Tassel-Baska and Brown (2007) suggest starting with the range of offerings in the local high school curriculum, such as honors and Advanced Placement classes, dual-enrollment, International Baccalaureate programs, art, music, theater, and the like. (If post-secondary schooling is available, then the range of options could be expanded to include these offerings as well.) This broad range of curricular options can be narrowed to the curricula at the middle-school level that would prepare students for or directly feed into these programs at the high-school level. Stepping back to the elementary level, schools can focus on talent development in those domains that feed into the middle school curriculum. Thus, the process starts with the diversity of talent development programs that are currently or potentially available at the high-school level and works backwards through the middle school and elementary years, at each step emphasizing development of those competencies that most directly serve as preparation for attainment at the next level. Good talent development programs offer multiple routes to engage and develop students' interests, creativity, and competence. The programs also have well-articulated links between curricula at different educational levels that help insure that students are not only having fun, but are also being prepared for further talent development (VanTassel-Baska & Brown, 2007).

## MOVING FROM TALENT TO APTITUDE

Although the term talent is ordinarily the preferred way to talk about giftedness with the public, an in-depth understanding of how best to develop talent requires more than simply identifying it. One must consider a much broader range of personal and environmental factors that can impact the process of turning talents into competencies (see, e.g., Gagne, 2009). Aptitude theory offers one approach. Aptitude is a word much like talent, but more inclusive. Aptitude includes the cognitive (or physical) characteristics that typically define the term “talent,” but it also includes any other personal characteristics that are required for successful learning (or performance) in a particular environment. Although usually grounded in biological predispositions, those characteristics that function as aptitudes are invariably developed through the child’s interactions with the environment. Some of these interactions are common to most children in a particular culture. Others are unique to the child’s family and its circumstances. And some are explicitly developed through schools and other social systems. For example, school achievements commonly function as aptitudes for future learning. Basic skills in reading, writing, and mathematics are required for the acquisition of expertise in all academic domains. Existing content knowledge in a domain provides the foundation for new learning, and thus current achievement in a domain functions as an important aptitude for new learning in that domain.

The word aptitude encompasses much more than cognitive constructs such as ability or achievement. Attaining competence in any domain of nontrivial complexity requires years of learning. Interest, motivation, and persistence are thus critical aptitudes for the developing expertise.

Finally, and most importantly, the term aptitude is not a descriptor of a person that is somehow independent of context or circumstance. Aptitude is inextricably linked to context. Indeed, defining the situation or context is part of defining the aptitude. Changing the context can subtly or substantially alter the personal characteristics that influence success.

Consider what transpires when students confront a new learning task. Every student comes to that situation with a repertoire of knowledge, skills, attitudes, values, motivations, and other propensities that they have developed and refined through life experiences to date. Situations sometimes demand, sometimes evoke, or sometimes merely afford the use of particular subsets of these characteristics. Of the many characteristics that might influence a person’s behavior in a particular situation, only a small set aides one in achieving his or her learning or performance goals. These characteristics function as aptitudes for that person in that situation. 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). Examples of characteristics that commonly function as academic aptitudes include the ability to understand and follow directions, to use previously acquired knowledge appropriately, to make good inferences and generalizations, to resist distractions, and to persist in the pursuit of excellence.

Not all of the characteristics that a person brings to a situation are helpful. Those characteristics that impede learning or performance function as inaptitudes. Common 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. Even characteristics that we would generally value might not be helpful in a particular context. For example, a teacher who values routine over innovation might interpret creativity as disruption. Many bright adults recall times when teachers valued being right more than encouraging students whose questions threatened their own feelings of competence. However, changing the context can replace rejection with acceptance and thereby transform the outcomes. Therefore, characteristics that functioned as inaptitudes in one context can be irrelevant or even function as aptitudes in a different context.

An aptitude theory of talent development thus helps dispel the common myth that superior general intellectual ability should be sufficient for the attainment of excellence. Many years ago, Thorndike (1963) pointed out that the concepts of over- and underachievement were grounded in this fallacy. The typical interpretation of underachievement, for example, assumes that the child has requisite aptitude for success, but for some reason fails to use that aptitude. A more reasonable interpretation,

Thorndike argued, is not underachievement, but rather that our model of what is required for learning is under-complicated. It is like assuming that all tall people should excel in basketball and calling those who do not excel “basketball underachievers.” Rather, the attainment of high levels of skill in basketball requires many other characteristics, both physical (strength, coordination) and psychological (interest in the game, competitiveness). Further, the context in which these skills must be developed defines the importance of some of these characteristics. Learning basketball on playgrounds in the inner city requires different personal characteristics than learning the game in the gyms of private schools.

### EFFECTS OF CONTEXT

The same situation that assists one student can thwart goal-attainment in another. For example, discovery-oriented or constructivist teaching methods generally succeed better with more able learners, while more didactic methods may work better with less able learners (Cronbach & Snow, 1977; Snow & Yalow, 1982). Less-structured learning situations allow more able students to use their superior reasoning abilities, which function as aptitudes. However, anxious students often perform poorly in relatively unstructured situations (Peterson, 1977). Thus, the same situation that affords the use of reasoning abilities can also evoke anxiety. Recent efforts to understand how individuals behave in academic contexts have emphasized the importance of interest, personality, and ability traits that commonly work together to produce the outcomes that we observe. For example, Ackerman (2003) found that high scores on measures of interest in developing social relationships had negative associations with knowledge acquisition, even in samples of talented individuals. This last finding reinforces the admonition that gifted students will vary as much from each other on those dimensions least correlated with *g* as students in the general population (Lubinski & Benbow, 2000).

Understanding which characteristics of individuals are likely to function as aptitudes begins with a careful examination of the demands and affordances of target tasks and the contexts in which they must be learned or 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 or makes likely or useful. Discovery learning often affords the use of reasoning abilities; direct instruction often does not.

Aptitude is linked to context. Unless we define the context clearly, we are left with distal measures that capture only some of the aptitudes needed for success. Because they lack context, *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. Averaging performance measures across learning situations and outcome measures often obscures the impact of the particular abilities and magnifies the relative importance of *g* (Lubinski, 2004; see Box 2). Domain-specific abilities add importantly to predicting school achievements in particular domains, even after controlling for the relationship between general ability and general school achievement (Gustafsson & Balke, 1993). Therefore, identifying talent must go beyond measures of general ability or academic competence to consider other personal and social skills that are required to succeed in the educational programs that are available—or that could be developed.

## MEASURING APTITUDES GENERAL PRINCIPLES

### TWO METHODS FOR INFERRING APTITUDE

Aptitude is commonly inferred in two ways. In the first way, aptitude for learning in a domain is estimated from the ease or speed with which the individual acquires competence in that domain. Aptitude is then inferred retrospectively, when the

individual learns concepts or skills after a few exposures that others learn only after much practice. When available, this information provides the most unambiguous evidence of aptitude. 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). However, people differ in their learning opportunities on the task itself or on other tasks that allow transfer to the target task, and so inferences about aptitude are defensible only when learning opportunities for the group being compared do not differ markedly. Tests include finely-differentiated age and grade norms because of the difference even a few months can make in the child's rank in cognitive or academic development. But such adjustments for opportunity to learn are effective only when the child has had learning experiences that are typical of others in the norm group. When a child has not had typical learning experiences, then even the most carefully developed national norms will over- or underestimate the student's aptitude. There are simple ways around this problem, but they require a rethinking of how to interpret scores on assessments.

In the second way, aptitude is inferred from performance on tasks outside of the domain of interest that require cognitive or affective processes similar to those required for learning in the domain (Carroll, 1974). Because these measures only predict success in the domain, they will more often err in identifying those students who will later excel than inferences based on performance to date in the domain itself. For example, dance instructors screen potential students by evaluating their body proportions, ability to turn their feet outwards, and ability to emulate physical movements (Subotnik & Jarvin, 2005). Although none of these characteristics require the performance of a dance routine, all are considered important aptitudes for acquiring dance skills. In the cognitive domain, measures of general ability are used because they measure thinking skills that are required for academic learning. For both dance and academic learning, these inferences are valid only if the students have had similar opportunities to develop the abilities required by the outside assessments.

Predictions about future learning based on current estimates of aptitude presume that the person, the content of the domain, and the learning environment will remain constant for future learning as for learning to date. In most cases, such a situation is only approximately true, especially as the time lag increases. Children develop new abilities, the learning demands change as children acquire expertise in a domain, and the learning environments themselves change as children change instructors, grades, and schools. If the child has had unusually strong or weak preparation for a particular school task, such as reading, then rank orders can change dramatically as other children catch up and sometimes surpass those who had early preparation. Status scores, such as percentile ranks (PRs) and IQs, reflect the consequences of these changes, but not the differential development that produced the changes in rank order. For these and other reasons, the identification of aptitude (or talent) must be an ongoing or at least regularly-repeated activity, not a one-time affair.

The primary aptitudes for academic learning are current academic knowledge and skills, reasoning abilities, interests, and persistence (Corno et al, 2002). All of these are best understood as tied to particular domains of learning and methods of instruction. For example, success in a verbally demanding domain, such as creative writing, requires verbal reasoning skills, knowledge of the conventions of the language and skill in applying them, interest in writing, and the ability to persist in the development of creative writing skills. Learning in other domains is similarly contextualized. Other chapters in the book discuss ratings scales, interest inventories, performance assessments, and other measures of aptitude. In this chapter, we focus on the measurement of reasoning abilities using individual and group intelligence tests. [Chapter 10](#) illustrates procedures for integrating some of these other measures into a talent identification system.

## HIGH-ACCOMPLISHMENT VERSUS HIGH-POTENTIAL STUDENTS

Two groups of children should be considered when designing programs for the academically gifted (Robinson, 2008). Although most students fall somewhere between the poles that these groups define, naming the endpoints of the continuum

helps clarify the children the programs should be serving.

The first group consists of those students who currently display academic excellence in a particular domain. We refer to these students as belonging to the high-accomplishment group. Although the measurement of accomplishment in any domain is not a trivial matter, these students are generally easier to identify than those in the second group. Students in the second group do not currently display excellence in the target academic domain, but are likely to do so if they are willing to put forth the effort required to achieve excellence and are given the proper educational assistance. We refer to these students as belonging to the high-potential group. Students commonly fall in the high-potential group because, through age, circumstance, or choice, they have not developed expertise in a particular domain. Because of their necessarily limited experience, young children most commonly fall in this group. Even the measures of accomplishment that we assess for young children, such as reading skills, are best seen as aptitudes that will later be needed for the acquisition of significant subject-matter expertise. Indeed, if one defines scholarly productivity or artistry in a domain as something beyond expertise (Subotnick & Jarvin, 2005), then even the most accomplished students will at best exhibit high potential. If, on the other hand, expertise is defined in terms of academic achievement well in advance of age or grade peers (e.g., Gagne, 2008), then many more children will exhibit high accomplishment. However, some students who do not display high accomplishment might also currently do so had they had the opportunities to develop these skills. Put differently, high-potential students display the aptitude to develop high levels of accomplishment offered by a particular class of instructional treatments.

High-accomplishment students typically need different educational programs than high-potential students. An undifferentiated label such as “gifted” does not usefully guide educational programming for a group that contains a mix of both high-accomplishment and high-potential students. Both groups need instruction that is geared to their current levels of accomplishment. Because their levels of accomplishment differ, instruction aimed at one group will often be inappropriate for the other group. (For example, high-accomplishment students often can best be served by academic acceleration within the domains in which they excel.)

The distinction between high-potential and high-accomplishment students is especially important in the identification of academically talented minority students. Many of the most talented minority students will not have had opportunities to develop high levels of accomplishment in the skills valued in formal schooling. Therefore, identifying such students depends on a clear understanding of how one can make inferences about academic aptitude using measures of both learning-to-date and potential for future learning. Providing advanced learning opportunities that meet the needs of both groups of students often requires differentiating the curriculum so that students are exposed to materials and opportunities specific to their individual needs.

## GATHERING DATA ON STUDENTS

### MEASURING DOMAIN KNOWLEDGE AND SKILLS

Academic learning at all levels builds on what the student already knows and can do (Glaser, 1992). Measures of current knowledge and skill are therefore usually the best predictors of future success in similar academic environments, especially when new learning depends heavily on old learning. The more closely measures of accomplishment sample critical aspects of emerging expertise in the domain, the better they will capture aptitude for learning in that domain. Measures of current knowledge and skill include on-grade-level and above-grade-level achievement tests, end-of-course examinations, and well-validated performance assessments such as rankings in debate contests, art exhibitions, and science fairs. Teacher grades and a host of other measures of academic learning are sometimes useful, but often have serious limitations because of their subjectivity or the limited sample of behavior that they represent. For example, even though trained raters can give highly

reliable ratings of a particular essay, performance on one essay shows only modest correlations with performance on another essay.

Although useful as a measure of basic skills in reading and numeracy, most achievement tests—especially those designed for children in the elementary grades—contain relatively little content knowledge, particularly in domains such as history, social science, literature, and the physical sciences. 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. But bright children assemble vast amounts of knowledge about specific topics that are, at best, represented only superficially on achievement tests. In this respect, the dilemma that confronts those who assess gifted children is the same dilemma that has stymied those who investigate adult intelligence. Adults develop substantial competencies in specialized domains. Any attempt to assess their functional intelligence by a test that can be administered to all adults misses much more than it captures. Hunt (2000) suggests that we might do a better job if the metaphor that guided assessment construction were to conduct an inventory rather than a survey. When designing educational interventions, this metaphor asks us to attend to what children know and can do in domains of their own choosing and the activities that fascinate them.

Out-of-level testing is a useful way to gather more information on the student's achievement than can be gleaned from an on-level-test that is too easy for the student. Tests such as PLAN and EXPLORE Global (published by ACT, Inc.) can give estimates of achievement, but because they are short, they include only a few items at each grade level. Above-level versions of longer, norm-referenced achievement tests, such as the Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2005), offer more specific information on levels of competence within domains. As children mature, end-of-course exams and exams designed for high-school students that are modeled after the Advanced Placement exams or the SAT II exams (published by the College Board) can provide useful information on the level of development within particular academic domains. The increasing availability of computer-based tests delivered over the internet should make it easier for school personnel to gather such information on students.

Whenever such tests are used to help make decisions about possible acceleration of a student within a domain, the test should be aligned with the local curriculum, and the student's performance on the test should be compared with the performance of students in the potential accelerated classroom. Norm-referenced achievement tests can be used in this way if the child is administered the same level of the test that is administered to students in the potential accelerated class. Sometimes other measures are available that offer even more focused estimates of readiness. For example, many schools in Iowa administer the Iowa Algebra Aptitude Test (IAAT; Schoen & Ansley, 2005) to all sixth-grade students. The performance of the child on this test can be compared with the performance of children in the school who will be enrolled in Algebra. A useful rule of thumb is that the child should score at about the 80th percentile (or higher) in distribution of scores for the target class before being accelerated into the class. Of course, many other factors should be considered before making such decisions as is discussed in the Iowa Acceleration Scale (IAS; Assouline, Colangelo, Lupkowski-Shoplik, Lipscomb, & Forstadt, 2009).

## MEASURING REASONING ABILITIES

Although current knowledge and skill in a domain are often the most important aptitudes for new learning in that domain, other aptitudes enter the picture with each step into the future. For example, given the same type of instruction, continued improvement in a domain requires interest, or at least persistence. More commonly, continued success requires a new mix of abilities: Algebra requires important thinking skills not needed in arithmetic; critical reading requires thinking 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, developing expertise



requires mastering new and, in some cases, qualitatively different types of tasks at different stages. Sometimes the critical factor is not only what is required for success, but also what is allowed or elicited by the new context that might create a stumbling block for the student. For example, in moving from a more structured to a less structured environment, a student may flounder because he is anxious or is unable to schedule his time.

It should be no surprise, then, that the second most important set of personal characteristics for academic learning is the ability to go beyond the information given, to make inferences and deductions, and to see patterns, rules, and instances of the familiar in the unfamiliar. The ability to reason well in the symbol system(s) used to communicate new knowledge in that domain is critical for success in learning. Academic learning relies heavily on reasoning (a) with words and about the concepts they signify and (b) with quantitative symbols and the concepts they signify. Thus, the critical reasoning abilities for all students (minority and majority, monolingual and multilingual) are verbal reasoning and quantitative reasoning. Nonverbal reasoning abilities are less important and show lower correlations with school achievement, especially in verbally demanding domains (Anastasi & Urbina, 1997; Gohm, Humphreys, & Yao, 1998; Lohman, 2005b; Thorndike & Hagen, 1987; 1995).

Good reasoning tests will present a broad sample of different reasoning tasks. Performance on a set of items that follow the same format provides only limited information about individual differences in broad ability constructs such as "reasoning." The term ability implies consistency in performance across a class of tasks that vary widely in their surface features. Psychological tests are simply organized collections of such tasks. However, typically less than half of the variation on well constructed, reliable tests is shared with other tests that measure the same construct using somewhat different kinds of tasks. An early, but still reasonable rule in psychological measurement is that when measuring any ability, one should combine performance across at least three different measures that use different formats to reduce the specific effects of individual tasks (Mulaik, 1972). This fundamental principle of psychological measurement is sometimes violated on individually administered ability tests that attempt to estimate many different abilities in a single testing session with fewer than three subtests and on group-administered tests in which all items follow the same format.

#### INDIVIDUALLY-ADMINISTERED INTELLIGENCE TESTS

Individually-administered intelligence tests are often used in the identification of academic talent and, in some cases, function as the primary criterion for admission to programs. Indeed, for many years, an IQ score of 130 on an individually-administered intelligence test was the standard requirement. Individually administered tests are now used much less frequently, primarily because of the cost and inconvenience of arranging for each child to be tested by a professional psychologist. However the requirement of an IQ of 130 is still widely used, albeit often disguised as a national percentile rank of 97 or higher.

#### EQUITY AND FAIRNESS ISSUES.

Individualized assessments for the purpose of identifying giftedness almost always must be paid for out of pocket, which raises issues of equity (Renzulli, 2005). Parents who have the financial resources to pay for testing are more likely to secure outside testing for their child than are parents without these advantages. Furthermore, scores from outside examiners may not be comparable with scores on group-administered tests that are administered to other children. For example, scores on the individually-administered test can be inflated because the child has been tested several times or because the examiner used a test with out-of-date norms. Or, scores on the individual test can be depressed because the examiner reports Full Scale scores rather than scores that exclude subtests that have lower *g* loadings. Therefore, if individual tests are required or accepted, policies should be developed that explicitly name which tests and test scores will be accepted and that inform parents of their

rights to obtain such testing if they cannot afford it.

Inferences about ability or talent require comparison of a child's performance on a test with the performance of other children who have had similar opportunities to develop the knowledge and skills required by the test. The assumption of similarity is often indefensible when national norms are used to interpret the performance of poor, minority, and ELL children, even on nonverbal tests. Subgroup norms that compare a student's performance on the test to the performance of other children who have had similar learning opportunities make the assumption more defensible. With the notable exception of the WISC-IV Spanish, such norms are not available for individually administered tests. However, unless all students are administered the test, even the WISC-IV Spanish cannot provide the local norms that are recommended for school-based talent development programs. Furthermore typically only children who are nominated for the program are tested individually, which can exclude children who do not conform to the teacher's conception of a gifted child. For these and other reasons, individually administered tests should not be the primary tool for screening admission to school-based programs.

### SOME APPROPRIATE USES.

Individually administered tests do have important uses for gifted identification and placement (Robinson, 2008). Admission to special programs for the exceptionally or profoundly gifted offers a clear example. When administered in conjunction with a battery of other assessments, individually administered ability tests also can assist in the identification of twice-exceptional students (i.e., students with a disability as well as a talent; Assouline, Foley Nicpon, & Bramer, 2006; Kaufman & Harrison, 1986; Newman, Sparrow, & Pfeiffer, 2005). Individual testing provides the opportunity for a professional psychologist to observe behavioral and neurological factors that could impair performance. For example, a child may perseverate unnecessarily on difficult problems, display high levels of anxiety, or exhibit other behaviors that impact performance on the test or in the classroom.

### INTERPRETING SCORE DISCREPANCIES.

Even when scores on individual ability tests are available and the norms are appropriate, interpretation of those scores is not straightforward. Gifted students show larger discrepancies between subtests than average-ability students (Saccuzzo, Johnson, & Russell, 1992; Sweetland, Reina, & Tatti, 2006; Wilkinson, 1993). Composite or full scale scores poorly summarize the abilities of examinees when the score profile has marked peaks and valleys. Areas of exceptional ability are masked by lower scores when the composite is computed.

Large discrepancies among subtests for highly able students also cannot be interpreted the same way similar discrepancies for average-ability students are interpreted. For example, a superior score on a verbal comprehension index coupled with a high average score on a working memory index does not necessarily mean anything clinically important (Newman, et al., 2008). Or, when verbal skills are significantly higher than nonverbal skills in a high-ability student, a clinician would need to gather corroborative evidence before considering a nonverbal learning disability diagnosis (Assouline, Foley Nicpon, & Whiteman, 2010).

### USING COMPOSITE SCORES FROM MODERN ABILITY TESTS.

There also has been much discussion about whether the composite or full scale scores on modern ability tests should be used for making inferences about giftedness. Modern intelligence tests sample a much broader range of abilities than older intelligence tests. The number of scores on the Stanford Binet increased from the single IQ score on the first edition to the five separate scores on the fifth edition of the test (Roid, 2003). Similarly, the two scores on the initial Wechsler tests doubled to

four separate scores on the fourth edition of the test. The Woodcock-Johnson III Tests of Cognitive Abilities (WJ III COG) offers even more scores (see Box 3).

The multiple abilities measured by modern intelligence tests are helpful for making diagnoses of learning problems. They can be less helpful for making inferences about academic giftedness. Students classified as gifted by other methods obtain much higher scores on tests that require reasoning than on the tests that emphasize working memory, perceptual speed, and other more specific abilities. For example, academically gifted students often obtain much higher scores on the Verbal Comprehension and Perceptual Reasoning indexes of the Wechsler Intelligence Scale for Children (WISC-IV) than on the Working Memory and Processing Speed indexes (Williams, Weiss, and Rolhus, 2003; Newman et al., 2008; Rimm, Gilman, & Silverman, 2008). Since all four of the WISC-IV indices contribute to the Full Scale IQ score, gifted students sometimes obtain lower than expected Full Scale IQ scores. In an effort to circumvent this problem, as well as minimize misinterpretation of the Full Scale IQ when the four factor indices are discrepant, a composite called the General Ability Index (GAI) was introduced. Only the Verbal Comprehension and Perceptual Reasoning indexes contribute to the GAI score. Therefore, examiners are encouraged to rely on the GAI rather than the Full Scale IQ when making inferences about giftedness (NAGC, 2008).

The admonition not to use the WISC-IV Full Scale score can be confusing to those who equate intelligence with  $g$  and thus define giftedness as a high level of  $g$ . An aptitude perspective on talent helps explain why narrower measures of ability, such as the old Stanford-Binet LM or the WISC-IV GAI, may better predict academic success in some domains than the full WISC-IV battery. As discussed, an aptitude perspective requires that one specify the psychological processes that are needed to learn and perform well in a particular class of situations. An undifferentiated construct such as  $g$  is not nearly as helpful for talent identification as a list of more specific constructs that capture the ability to reason in the symbol systems used for developing expertise and communicating new knowledge in a particular domain. When stated in this way, it is clear that the sort of abstract, verbal reasoning abilities measured by the older forms of the Stanford-Binet will always be critical for many forms of academic learning. But this formulation also makes it clear that reasoning abilities in other symbol systems (e.g., quantitative, spatial, musical) will be even more important than verbal reasoning for talent development in quantitative, spatial, and musical domains.

### SCALING ISSUES.

Another debate about individual tests is whether the procedures used for scaling modern intelligence tests give scores that properly reflect the abilities of profoundly gifted individuals. For many years, degrees of giftedness followed the classification scheme based on IQ scores on the 1960 revision of the Stanford-Binet.

Moderately Gifted 125+

Highly Gifted 145+

Exceptionally Gifted 160+

Profoundly Gifted 180+

Modern tests do not produce such high scores, in spite of heroic efforts to provide extended norms for both the Stanford Binet, Fifth Edition (SB-5) and the WISC-IV (Roid, 2003; Zhu, Clayton, Weiss, & Gabel, 2008). For this reason, some have argued that the old Stanford-Binet LM (SB-LM) is a better test for distinguishing among exceptionally and profoundly gifted children (Silverman, 2009). That newer tests give fewer extremely high scores is not a consequence of including subtests for specific abilities such as working memory and perceptual speed. If that were the case, then the new tests would identify the same number of high-scoring students as the old SB-LM identified when its norms were current. Rather, the old and new tests

would simply identify somewhat different groups of students. It is important to understand why this is not the case.

The standardization sample for most ability tests includes only about 100 individuals in each age group. Thus, on average, there will be only one examinee that scores in the 99th percentile of the distribution in each age group. Even if the sample included 1000 individuals at each age, there would be only 10 individuals with IQs of 135 or greater. Given this paucity of individuals at the tails of the score distributions, normative standards for extreme scores either implicitly or explicitly rely on some model for how scores are distributed in the population. The most common assumption is that scores should follow a normal distribution just as is observed for many other characteristics that we can measure with great confidence. The assumption that the distribution of ability for the non-clinical population is approximately normal is probably more defensible than that it is not normal. It is surely more convenient. The debate could be resolved if we had objective scales for ability and then could empirically determine if the observed score distributions were in fact normal. Modern ability tests (like the SB-5 and WJIII COG) that are constructed using only those items that fit an Item Response Theory (IRT) model come closer to the ideal of an objective score scale than older tests. Empirical distributions of IRT-based scale scores for the very large CogAT standardization samples do in fact appear to be normal (see Lohman, 2010, for a non-technical summary of the norming process), which strengthens the assumption of normality. Unlike CogAT Standard Age Scores, however, deviation IQ-like scores on modern tests are based on much smaller samples and thus are constructed by assuming that the scale scores are distributed normally within all age groups.

IQ scores on the SB-LM were computed by forming the ratio of the scale score or Mental Age (MA) on the test by their chronological age (CA). The assumption that the IQ distribution is not normal is inherent in the computation of ratio IQs, even the modified ratio IQs used in the SB-LM. Growth in MA is largest at the early years and gets smaller until leveling off at about age 16. The leveling off of MA was handled by fixing the maximum CA at 16. However, the variability of the distributions of ratio IQ's differed across ages. When Form LM was standardized, these variations in standard deviation across age were removed by multiplying the observed standard deviation of scores at each age by whatever constant that made the standard deviation equal to 16. However, adjusting or modifying IQs in this way did not remove skewness in the score distributions.

IQ distribution gets skewed to the right when the denominator (i.e., CA) has a much smaller range than the numerator (i.e., MA). In such cases, small differences in mental age translate into very large differences in the corresponding ratio IQ score. For example, consider the 6-year-old child who has a mental age of 9, or three years in advance of age-mates. The corresponding ratio IQ is 150. Now consider the same child with a chronological age of 9. If she is still mentally three years in advance of peers, her MA will be 12 and the ratio IQ only 133. In order to preserve the 150 IQ, her mental age would have to be 13.5, or 4.5 years in advance of her peers. This is why exceedingly high ratio IQ's become increasingly unlikely as the denominator (CA) gets larger. It is also why "very few clinicians have been able to make practical use of the [Stanford Binet-LM] for bright children over the age of 10" (Ruf, 2003, p. 5).

The spreading out of scores for young children at the extremes of the ratio IQ scale is viewed as a positive attribute of the SB-LM by clinicians who want to distinguish among the highly and profoundly gifted (Silverman, 2009). Although spreading out the test scores in this way may be helpful, the corresponding normative scores (i.e., IQs) cannot be trusted both because they are based on out-of-date norms and because the spread of IQ scores is a necessary consequence of the way ratio IQs are constructed, not a fact of nature.

### SHORT FORMS.

Short forms of individually-administered ability tests are popular among those who either cannot or do not wish to

administer a complete test battery. Examples include the Kaufman Brief Intelligence Test II (K-BIT-II) and the Screening Assessment for Gifted Elementary and Middle School Students, 2nd edition (SAGES-2). In addition to brevity, such tests have the advantage of placing the examiner in direct contact with the student.

There are several problems with these tests and similar brief ability estimates on other tests, such as the WISC-IV GAI score or the Brief Ability Index on the WJIII. First, because they are short, these tests (and indices) sample only a limited portion of a much larger domain. Broad constructs (such as fluid reasoning ability) are often under-represented. For example, the Reasoning subtest of the SAGES-2 uses only about 30 pictorial/figural analogy items. Similarly, the K-BIT-II requires only 15-30 minutes to administer both the Verbal (Riddles and Verbal Knowledge) and Nonverbal (Matrices) subtests. Second, because of the limited sampling of abilities, these tests cannot support inferences about ability profiles, which are needed for placement into programs that emphasize the development of different talents, such as mathematics versus literacy.. Third, scores on short tests are less reliable than scores on otherwise similar tests that present more items and use a broader selection of item formats. Longer tests are critical for identifying giftedness. As discussed below, errors of measurement are much larger for extreme scores than for scores near the mean. Short tests only compound this problem. Life-altering decisions about a child should be based on the best evidence that can be obtained. The few minutes saved by using short tests and indices come at a considerable cost to the children and the program.

Sometimes using a short test can be defended if it can successfully reduce the number of students who must be administered a more comprehensive assessment battery. But this practice must be done without screening out students who would have been admitted had they been administered the more comprehensive assessment. Unfortunately, this is not easy to do. It helps if the screening test is highly reliable, if items on the screening test sample from several different domains rather than a single domain (e.g., only figural reasoning), and if the cut score on the screening test is set quite low. More specific recommendations on how to implement a two-stage testing program that starts with a screening test are provided in [Chapter 10](#).

#### GROUP-ADMINISTERED ABILITY TEST

Group ability tests are commonly viewed as rough screening tools 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. Although such skepticism about group ability tests is amply warranted, as with most stereotypes, it is not always the case. The generalization is most likely to hold when the group-administered test is relatively short, samples only a portion of the cognitive domain, or has out-of-date or poorly developed norms. However, a group-administered test will sometimes give the better estimate of the student's academic aptitude than the individually administered test. This will happen if the group administered test samples more comprehensively from the domain of abstract reasoning abilities than the individually administered test.

Although many different tasks have been used to measure reasoning on group-administered tests, a few are used much more common than others: analogies, matrix problems, series completions, and classification tasks. Some test batteries also measure verbal reasoning through sentence-completion tests, sentence-comprehension tests, and even vocabulary. Others include more specific spatial tasks, such as form boards or paper-folding tests. And others use quantitative tests that require examinees to make relational judgments (such as greater than or less than) between quantitative concepts or to determine how numbers and mathematical operators can be combined to generate a product.

Some of the major group-administered ability tests are summarized in Box 4. The tests vary in the number and type of abilities they measure, the quality and recency of their norms, the support materials for teachers and parents, and their reliability.

Examples of the nine reasoning tasks used in the most recent version of Thorndike and Hagen's Cognitive Abilities Test (Form 7 of CogAT, Lohman, 2011) are shown in Figure 1 (on page xx). The left column shows an example of one of the language-reduced item formats that are used with younger children. The right column shows an example of the standard item formats used with older children. Each reasoning ability is estimated by three subtests that require somewhat different processing. Although each of the three batteries can be administered and interpreted independently, most users administer all three batteries. Those who need a shorter test can administer the Screening Test that contains only the first subtest in each battery. The three reasoning abilities measured by the complete test correspond with the three aspects of fluid reasoning ability identified in Carroll's (1993) compendium.

Because they measure a much broader sample of different abilities, individually administered ability tests developed in recent years actually have fewer and shorter reasoning subtests (Frazier & Youngstrom, 2007). Thus, the CogAT scores are generally more reliable (see Table 1 on page xx), a better measure of *g* (Lohman, 2003a, b), and equally good or better predictor of success in school.

### THE IMPORTANCE OF ABILITY PROFILES IN TALENT IDENTIFICATION

Although many have advocated for multidimensional theories of giftedness (Feldhusen & Jarwan, 2000; Gagné, 2009; Sternberg, 2003), children often are admitted to programs for the gifted and talented using group or individually administered measures of general ability (Assouline, 2003). Some programs require an IQ score of at least 130 (or national PR of 97). Other programs follow the recommended practice of collecting multiple sources of information, but unwittingly collapse it into a measure of general ability. For example, schools collect scores from different tests, classroom grades, and teacher ratings, assign points to each (often using somewhat arbitrary rules), and then add them together. Admission is based on the total number of points, which estimates the general factor common to all of the measures.

A single, omnibus score—whether from an intelligence test, an achievement test, or a more complex collection of ratings and assessments—will identify only that fraction of children who excel across all of the abilities measured by the tests and rating scales. The score profile for such students shows no unusual peaks or valleys. Only a minority of gifted students show this sort of flat score profile (Achter, Benbow, & Lubinski, 1997; Lohman, Gambrell, & Lakin, 2008). Indeed, as we will show below, the probability that a student has a flat score profile decreases as the students' scores depart from the mean. Furthermore, these differences in abilities have important long-term consequences. For example, profiles on ability tests administered at age 13 help predict the undergraduate majors students choose and the advanced degrees that they obtain (Lubinski, Webb, Morelock, & Benbow, 2001; Park, Lubinski, & Benbow, 2007).

### WHY PROFILES ARE OFTEN UNRELIABLE

Although some researchers have demonstrated that ability profiles have utility for both research and guidance counseling with gifted students, others have questioned both their reliability and usefulness. The criticisms have been especially strong for subtest scores on individually administered ability tests (Watkins, Gluting, & Youngstrom, 2005). Any effort to move beyond IQ's or similar omnibus scores assumes that the variations in scores captured by the profile reflect more than noise.

The simplest profile consists of scores from two tests. Even if the two scores are moderately reliable, difference between them can be quite unreliable. The difference score retains the measurement errors in both tests, thereby doubling the measurement error in the difference score from what it was in either of the original test scores. To make matters worse, the difference score discards any systematic variability that the tests share. The shared variance is reflected in the correlation

between the two tests. For example, suppose that the profile consists of scores on two, twenty-item tests. Because they are short, each test has only moderate reliability (say  $r_{xx'} = .75$  for both). Further, as usually is the case in ability testing, the tests correlate  $r = .6$  with each other. In this case, the reliability of the difference score (d) will be only  $r_{dd'} = .37$ .

The interpretation of a profile depends on which scores differ and by how much. The unreliability of these difference scores is what renders suspect most score profiles. Large differences between scores may not replicate and, conversely, seemingly similar scores may mask true differences. Thus, even if a student has equal abilities on several tests, it is likely that her scores will differ across the tests if the tests are short and therefore unreliable.

Although many factors influence the reliability of the difference scores that provide the unique information in the score profile, most important are the magnitude of the correlations among the separate scores and the reliability of each. Therefore, the first way to increase the reliability of the profile is to measure traits that are uncorrelated. Unfortunately, one can not increase reliability in this way when measuring abilities, especially abilities that require reasoning or problem-solving. Such abilities are strongly correlated.

The second way to increase the reliability of the profile is to increase the reliability of the separate test scores used in the profile. Reliability of test scores can be improved by increasing the number of items on the test. For example, if the lengths of two tests in the previous example were increased from 20 to 60 items, then reliability of each test would increase to  $r_{xx'} = .90$  and the reliability of the difference score would increase to  $r_{dd'} = .75$  (which was the reliability of the original, 20-item tests.) Longer tests make a big difference. As a rule of thumb, each test score in a profile should be based on at least 50 items in order to obtain adequate reliability to compare them. Unfortunately, administering tests with at least 50 items in each of several ability domains conflicts with most teachers and school administrators' desire to spend no more than a few minutes on ability testing. Unless those responsible for identification understand the problems created when using short tests, convenience will win the day and life-altering decisions will be made about children on the basis of inadequate and unreliable information.

Obtaining highly reliable subtest scores is difficult not only because the test battery often must be administered in a relatively short time, but also because it must measure the ever-growing list of abilities that characterize modern theories of intelligence (Frazier & Youngstrom, 2007). The number of scores on ability tests has increased from the single IQ score on the original Stanford Binet to four or five separate scores on the test's fifth edition (Roid, 2003). One cannot obtain reliable measures of four or five dimensions in approximately the same time that it took to measure one dimension.

One way to avoid this problem is to measure only those abilities that directly inform the identification process and to measure each as reliably as possible with several different subtests. Thorndike and Hagen (1971) followed this strategy when developing the CogAT. Rather than measure many different abilities, they measured only those reasoning abilities with predictive validity for educational achievement: verbal reasoning, quantitative reasoning, and nonverbal/spatial reasoning. Importantly, each of the three batteries of reasoning tests contained 60 to 65 items. Thus, although the correlations between scores on the batteries are high ( $r \approx .73$ ), the reliabilities of the batteries are considerably higher ( $r_{xx'} \approx .94$ ). From one-third to one-half of the reliable variance on each battery is not shared with one of the other two test batteries, thus allowing meaningful interpretation of the information that is embedded in the test profiles.

#### CAPTURING THE UNIQUE INFORMATION IN SCORE PROFILES

Every score profile contains three kinds of information: altitude, scatter, and shape (Cronbach & Gleser, 1953). Altitude refers to the overall height or level of the score profile. It reflects the influence of the general factor common to the scores in the profile. It is estimated by the average of the separate scores. Scatter refers to the variability of subtest scores. Finally, shape refers to the particular pattern of elevations and depressions in the profile. Profiles with the same amount of scatter can have

quite different shapes.

The shape of the profile can be indexed by comparing the similarity of the score profile with some standard. A logical standard can be established by enumerating the patterns that could logically be observed. This option is reasonable when the number of test scores in the profile is relatively small and the correlations among them are fairly uniform. For example, for the three CogAT scores [Verbal (V), Quantitative (Q), and Nonverbal (N)], there are three profiles that show a strength on one battery (V+, Q+, or N+), three that show a weakness (V-, Q-, or N-), and six more that show both a strength and a weakness (V+Q-, V+N-, Q+N-, Q+V-, N+V-, and N+Q-). All non-flat score profiles can be classified into one of these 12 categories.

#### FREQUENCY OF OCCURRENCE OF DIFFERENT SCORE PROFILES

Are some profiles more common than other profiles for the most or least able students? Figure 2 shows the percentage of students in the 2000 CogAT US standardization sample who had different score profiles on the CogAT multilevel battery. Scores for the Verbal, Quantitative, and Nonverbal Reasoning batteries are highly correlated, indicating a strong general factor that is virtually coincident with the general factor on the WISC-III (Lohman, 2003a). A strong general factor suggests that score profiles would not be dependable. As it turns out, this is not the case.

Figure 2 shows that about 40% of the students showed an approximately flat profile for Verbal, Quantitative, and Nonverbal reasoning. The performance of these students is well summarized in a single composite or average score. However, the majority of students (60%) showed a profile in which scores differed significantly. A significant or extreme strength [a score that was 20 or more Standard Age Score (SAS) points higher than the other two scores] was most common among students with a median stanine score of 1 and least common for students with a median stanine score of 9. A significant or extreme weakness (20 or more SAS points lower) shows the opposite pattern. The most able students were much more likely to show this profile than other students. Importantly, although only 3.4% of the population had profiles that showed an extreme weakness, 15.4% of the most able students showed this profile. Thus, profiles that show extreme differences between scores are much more common for the most and least able students than for average ability students. Consequently, one should never use the composite score on a diverse collection of tests when screening for giftedness. In addition, screening students on only nonverbal reasoning is also not advised. Most of the gifted students who obtain higher scores in quantitative reasoning or verbal reasoning than in nonverbal reasoning will not be identified if only those with high scores on the nonverbal test are given further consideration.

A moment's reflection shows why the frequency of extreme profiles varies by ability. For low-scoring students, an even lower score on the third battery is much less likely than a somewhat higher score. There is little room to move down and much room to move up. Conversely, high scoring students are much more likely to show a relative weakness than to show a relative strength on the third battery. This is what is meant, but not explained, by "regression to the mean."

#### REGRESSION TO THE MEAN

Any presentation of ability testing and giftedness would be incomplete without a discussion of "regression to the mean." Many in the field of special education ignore and some even deny the existence of regression to the mean. It is particularly odd that people trained in gifted education would do this. Regression to the mean, which was first observed by Galton when he plotted the heights of fathers and their sons, is one of the oldest empirical phenomena in the study of individual differences. It should be on page one of every text on gifted children. Regression is not acknowledged because many hold entity beliefs about abilities. Such beliefs are encouraged by the use of labels such as "gifted," by the use of status scores such as IQ or percentile rank that mask growth and other changes in ability over time, and by the human tendency to attend to information that



confirms our expectations and to ignore information that challenges our beliefs (Lohman & Korb, 2006).

Regression to the mean will be observed whenever the correlation between two variables is less than perfect ( $r < 1.0$ ). Anything that reduces the correlation between two sets of observations increases the amount of regression that will be observed. Some of this regression reflects errors of measurement. Measurement errors include not only the random fluctuations in attention or memory that impact thinking on individual items on a particular occasion, but also the selection of items on the test and format of the test itself. To make an inference about a child's level of reasoning ability, one must go beyond the child's performance on a particular set of test items that were administered on a particular day. The best estimates of error include changes in performance across time, forms of the test, and even test formats. Internal consistency estimates of reliability do not capture any of these sources of error. Estimating reliability by administering parallel forms of the same test on different occasions comes closer. Administering a different test on each occasion is best. For example, internal consistency estimates of reliability for the WISC-IV Full Scale scores are about .97. However, the correlations between WISC-IV Full Scale scores and total scores on other ability tests (administered within a one month period) are in the .6 to .8 range (Daniel, 2000; Wechsler, 2003).

### ESTIMATING EXPECTED REGRESSION

The expected regression in scores is easily estimated from the correlation. The predicted score on test 2 is simply

$$\hat{z}_2 = z_1 \times r_{12}$$

where  $\hat{z}_2$  is the predicted standard score on test 2,  $z_1$  is the standard score on test 1, and  $r_{12}$  is the correlation between the tests 1 and 2.

The expected score at time 2 will equal the score at time 1 only if the correlation is 1.0 or if the standard score at time 1 is zero (i.e., the mean). The lower the correlation is, the greater the expected regression. Indeed, when the correlation between two tests is zero, the expected test score at time 2 is the mean (i.e., 0) for all test takers. Although there is no regression at the mean (i.e.,  $z_1 = z_2 = 0$ ), the amount of regression increases as scores depart from the mean. Students who receive extremely high scores on test 1 are unlikely to receive similarly high scores on test 2. Box 5 gives some examples.

Regression will also be observed when two tests measure different constructs. Consider the three CogAT batteries. The reliabilities of scores on the three batteries are much higher than their correlations with each other (see Lohman, Gambrell, & Lakin, 2008). Therefore, the largest contributor to regression in the scores obtained on two different batteries is not error of measurement, but the fact that each test battery measures somewhat different abilities. The probability of obtaining an extremely high score on all three test batteries is much less than the probability of obtaining a high score on one battery. On CogAT, four out of every 100 students in grades 3 – 12 obtain a stanine score of 9 on each battery. However, only 2.4 out of every hundred obtain a stanine score of 9 on any two batteries. And only 4 out of every 1000 obtain a stanine score of 9 on all three batteries. Selection rules that admit only those who obtain high scores on multiple tests admit only a very small and unusual set of students (see chapter 10).

The greater frequency of one extremely low score for high-altitude profiles is not unique to CogAT. Indeed, the lower the correlation among scores in the profile (for reasons including multidimensionality as well as measurement error), the more common it will be. As a result, the CogAT authors have advised test users not to use the three-battery composite score to screen children for admission to programs for the gifted (Lohman & Hagen, 2001b; Thorndike & Hagen, 1986; 1994). Many quite capable students are likely to have one battery score that is low enough to bring down their averaged composite score. Rather than using only an estimate of  $g$ , the better procedure is to match the selection criteria with the demands of the

educational program (Lohman, 2005a; Renzulli, 2005). At the very least, schools should distinguish between verbal and quantitative/spatial abilities when identifying academic talent.

### CAUTIONS AND CLARIFICATIONS.

Regression is sometimes not observed (especially in the average scores of a group of students) because it is counteracted by practice or training effects, which can be substantial (Kaufman, 1994). Such effects are especially likely for students who are unfamiliar with the test formats and on tests that are most susceptible to practice effects, such as Block Design from the WISC-IV, matrix items on nonverbal tests, or any test that is speeded.

Although errors of measurement (broadly construed) are the largest contributor to regression for tests measuring the same construct, even error-free measures show regression, especially as the time lag between the two assessments increases. Cognitive abilities develop at different rates in children and the sources of individual differences change as one moves up the developmental scale. (See Lohman & Korb, 2006 for discussion). Regression to the mean is about change in rank order, not about the change in absolute score levels. Even if all of the students in an educational program improve, regression will be observed if all do not improve equally. The critical mistake is to assume that the score on an ability test reflects a fixed characteristic rather than relative status on a fallible estimate of a constantly growing (and changing) set of mental characteristics. Indeed, in order to maintain a particular rank across time, a child must not only get better each year but must improve at the same rate as others who had the same initial rank.

### ERRORS OF MEASUREMENT

Of the many things that one might want to know about the scores on a test, group or individually administered, 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 of test scores. The SEM makes it much easier to estimate the magnitude of score changes one is likely to see upon retest, at least for students who do not have extreme scores. For example, consider three ability tests ( $M= 100$ ,  $SD = 16$  that have reliabilities of .80, .90, and .95. The corresponding SEM's are 3.2, 1.6, and .8 IQ points. Put differently, changing the reliability from .8 to .9 cuts the expected error in half. Going from .90 to .95 halves it again. What are the consequences for a 95 percent confidence interval around the observed IQ score? The respective intervals have widths of 25, 12.4, and 3.1 IQ points. The confidence interval for the test with a reliability of  $r_{xx'} = .8$  is 8 times larger than the confidence interval for the test with a reliability of  $r_{xx'} = .95$ .

Table 1 shows how errors of measurement vary across several widely used individually- and group administered ability tests. For ease of comparison, all scores are reported on a scale with mean 100 and standard deviation 16. (see also chapter \_\_.)

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 SB-V) and between tests. In general, reliability increases with the number of items. For example, each level of the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997) has 38 items. Students are allowed 30 minutes to attempt as many as they can. It has the largest SEM of any test in Table 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 percent confidence interval for a student who receives a score of 100 on the NNAT is 88 to 112 – a range of 24 points. For the CogAT Composite score it is 95.6 to 104.4 – a range of 8.8 points.

### CONDITIONAL SEMS.

The concerns associated with SEMs are actually substantially worse for scores at the extremes of the distribution, especially when scores approach the maximum possible on a test, such as when, on a group test, students answer most of the items correctly. In these cases, errors of measurement for scale scores will 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 (Lord, 1980). This increase in errors of measurement for scale scores (but not for number correct scores) is shown in Figure 3 for Level A of the Verbal Battery of CogAT– Form 6. The errors increase for Universal Scale Scores (USS) because transforming raw scores into scale scores stretches the score scale at the extremes of the distributions. This increase in error for scale scores 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 (PR) scale because PRs are compressed at the tails of the distribution, whereas scale scores are spread out. For example, for tests such as CogAT and OLSAT that have a mean of 100 and standard deviation of 16, every Standard Age Score (SAS) above 134 receives the same PR of 99.

These relationships are shown in Figure 4. Notice that we have extended the upper scale so that the maximum SAS score is 150. Then, immediately below the figure, we have shown the raw scores that correspond to this SAS score for a 10-year-old child taking the CogAT Verbal Battery. Notice that once we get above 130 every additional item adds 10 SAS points. As a result, the error of measurement is much larger for high scores than for scores at, or, in this case, below the mean.

#### OUT-OF-LEVEL TESTING.

The best 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 to make 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 a 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 percentile are advised to move up two test levels (Lohman & Hagen, 2002), which is easy to do because all tests at these grades have the same directions and time limits.

People who balk at the need for this sort of adjustment must realize that it is precisely what happens when the child is tested individually. The examiner presents items until the student makes several mistakes in a row. Group tests also contain a single, long set of items for each subtest. However, for the convenience of administering the same test to an entire class, only a portion of the items are included in the common test booklet for each test level. The test constructor divides the long set of items into overlapping sets of items that progressively increase in difficulty. However, only those items that can be administered to the typical class will be included in the test. Children who are 3 to 4 standard deviations above the mean need items that are appropriate for children several years older than the typical student of the same age. It is helpful to remember that these will be the same children who will be earning SAT scores of 500 or above when they are 12 or 13 years old.

Because they are more easily adapted to the ability of the student, scores on individually administered tests do not show such large increase in SEM for high-scoring children. Nonetheless, even on these tests, true scores (i.e., the scores 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).

#### CONTROVERSIES ABOUT NONVERBAL REASONING TESTS

Nonverbal tasks like the CogAT Figure Matrices and Figure Classification tests shown in Figure 1 (on page xx) have long

formed an important part of both individual intelligence tests and group ability tests. Scores on the nonverbal batteries of these tests provided one indicator of ability for native speakers of the language, but often served as the only measure of ability for examinees who were not fluent speakers of the language. However, measurement experts have long cautioned that nonverbal reasoning tests do not capture the same ability construct that is measured by tests that use language (Anastasi, 1937) and therefore should not be used alone to make decisions about academic giftedness (Terman, 1930) or general intellectual competence (J. Raven, Raven, & Court, 1998; McCallum, Bracken, & Wasserman, 2001; see Box 6). Language, mathematics, music, and art are not contaminants grafted onto a fixed, innate intelligence, but rather critical vehicles for the development and expression of intelligence in particular symbol systems. Paring the world down to the small set of geometric forms used on a nonverbal reasoning test carves off much that any reasonable person would call intelligent thought.

Even though most nonverbal reasoning tests such as Raven's Progressive Matrices Test (Raven et al., 2000) and the Nonverbal Battery of the CogAT are reasonably good measures of *g*, they do not measure the verbal and symbolic reasoning abilities that are required for academic success for students from all ethnic backgrounds (Gustafsson & Balke, 1993; Keith, 1999; Lohman, 2005b). Even on the longest and most reliable nonverbal tests, only about half of the variance in test scores can be explained by *g*. The remaining variance is explained by other abilities, task-specific factors, and errors of measurement, which means that individual differences in the scores that students actually obtain on these tests are as likely to reflect factors other than *g* as they are to reflect *g*. Somewhat surprisingly, spatial abilities are only a small part of this non-*g* variance. Other tests that specifically measure spatial ability are needed to identify students who excel in visual thinking (Lohman, 1994).

#### IDENTIFYING ACADEMICALLY TALENTED ELL CHILDREN.

The main reason that schools use nonverbal tests, however, is not because they hope to identify visual-spatial learners or even because they believe that nonverbal tests are a good way to measure academic giftedness. Rather, the overriding reason is that differences between native and non-native speakers of English are substantially smaller on these tests than on tests that use English. For example, Lohman, Korb, and Lakin (2008) found that differences between the mean scores of ELL and non-ELL students were half as large on the CogAT Nonverbal Battery as on the CogAT Verbal Battery. There is thus a tradeoff in using nonverbal tests. On the one hand, nonverbal reasoning tests can reduce the amount of construct-irrelevant variance in test scores for nonnative speakers by reducing the impact of language. But by not measuring the ability to reason using verbal or quantitative symbol systems, nonverbal tests seriously under-represent the construct of academic aptitude (Braden, 2000; Lohman, 2005b; Mills & Tissot, 1995).

This loss in validity is shown not in the mean or average scores of groups, but in the correlations between test scores and various criteria of academic success. Unlike the means, the magnitude of these correlations does not differ across ethnic groups. Correlations between nonverbal, figural reasoning abilities and reading achievement typically range from  $r = .3$  to  $r = .5$ ; correlations with mathematics achievement typically range from  $r = .4$  to  $r = .6$  (Lohman & Hagen, 2002; Naglieri & Ronning, 2000b; Powers, Barkan & Jones, 1986). Although significant, these correlations are considerably smaller than the correlation of  $r = .8$  between verbal reasoning and reading achievement or between quantitative reasoning and mathematics achievement (Lohman & Hagen, 2002; Thorndike & Hagen, 1995). Lower predictive validity substantially impairs the ability of the test to identify academically talented students, regardless of ethnicity or language background. Put another way, using the nonverbal test will admit more ELL students, but will miss many of the most academically talented ELL students.

#### THE IMPORTANCE OF ACCOUNTING FOR OPPORTUNITY TO LEARN

The fact that the predictors of academic success are the same for all ethnic groups means that the identification of

academic talent requires measurement of the same aptitude variables for all children. What it does not mean is that the test scores of all children should be compared to a common norm group. Rather, inferences about talent (or aptitude) require the simple step of comparing a child's performance to that of other children who have had roughly similarly opportunities to develop the abilities measured by the test. This is not an option, as it is when making inferences about achievement. The statement that a 7-year-old child is reading English-language texts at the third grade level is meaningful regardless of the child's familiarity with the English language. However, making an inference about ability or talent from these same test scores requires first controlling for the child's opportunity to learn the English language. That this is not commonly done says more about naïve beliefs about what ability tests measure and, historically, of the difficulty of controlling for anything beyond age (in years and months) or grade (also in years or months). Norms tables for ability tests that make careful adjustments in estimated ability from small changes in age assume that the number of years and months since the child's birth provide a reasonable estimate of opportunity to learn. But this is not the case whenever the child's experiences differ markedly from those of other children in the norming sample. Simply including some of these children in the norming sample does not fix the problem. With the advent of personal computers, however, the task of separating test scores for all ELL and all non-ELL children is no more difficult than separating the scores for boys and girls or for third graders and fourth graders. The tradeoff is between obtaining precise estimates of talent using the wrong norm group or less precise estimates using a much better norm group. We demonstrate procedures for obtaining within-group comparisons in [Chapter 10](#).

ELL students who might someday excel as writers, mathematicians, or artists will generally show rapid learning when given the opportunity to learn concepts and skills in those domains. These students will also obtain higher scores on the verbal, quantitative, or spatial tests that measure the specific aptitudes required to develop competence in each domain than other ELL students (Corno et al., 2002). Measuring only nonverbal reasoning and not these more proximal aptitudes actually excludes most minority children who are likely to excel in domains that require more than nonverbal reasoning. But the development of these children will not be considered unusual unless their test scores are compared to the test scores of other children who have had roughly similar opportunities to develop the abilities that are measured (Lohman & Lakin, 2007). This applies to all abilities—even those abilities measured by nonverbal reasoning tests.

#### DO NONVERBAL TESTS LEVEL THE PLAYING FIELD FOR ELL CHILDREN?

It is commonly assumed and sometimes asserted that nonverbal tests level the playing field for ELL children (Naglieri, Booth, & Winsler, 2004). That this is not the case was shown in a large study in which approximately 2000 children (approximately 40% ELL) in grades K-6 were administered the Standard Progressive Matrices, the NNAT, and the CogAT by trained examiners. Directions were given in English or Spanish as appropriate. Large differences between the mean scores of all ELL and non-ELL children were reduced only slightly when the researchers controlled for ethnicity by comparing only Hispanic ELL and non-ELL students, all of whom were receiving free or reduced-price lunch at school. Non-ELL Hispanic students outperformed their ELL Hispanic classmates by 7.5, 7.3, and 9.5 IQ-like points on the Raven, CogAT, and NNAT, respectively. The data also showed that the norms on the Raven were about 10 IQ-like points too easy and that the norms on the NNAT were incorrectly computed, vastly over-identifying the number of high-scoring children.

Finally, as in several other studies (e.g., Carmen & Taylor, 2010) there was no support for Naglieri and Ford's (2003) claim that the NNAT identified equal proportions of White, Hispanic, and Black children. Indeed, ELL children were much more likely to receive very low scores on NNAT than on the other two nonverbal tests. Some of these findings are summarized in Figure 5.

Both naïve theories about what ability tests measure (Lohman, 2006a) and exaggerated claims for the efficacy of

improperly normed nonverbal tests (see Carman, 2010; Lohman, Korb, & Lakin, 2008) have misled many well-meaning educators. Nonverbal tests need not fulfill a utopian vision as measures of innate ability unencumbered by culture, education, or experience in order to play a useful role in the identification of academically gifted children. Nonverbal reasoning tests can help identify bright children, especially those who are poor or who are not fluent in the language of the dominant culture. When combined with measures of quantitative reasoning and spatial ability, nonverbal reasoning tests are particularly helpful in identifying students who will excel in engineering, mathematics, and related fields (Shea, Lubinski, & Benbow, 2001). But the development of these children will not be considered unusual unless their test scores are compared to the test scores of other children who have had roughly similar opportunities to develop the abilities that are measured by the test – even if it is a nonverbal reasoning test (Lohman & Lakin, 2007).

#### “NONVERBAL” MEASURES OF VERBAL AND QUANTITATIVE REASONING

In their book on nonverbal testing, McCallum, Bracken, & Wasserman (2001) distinguish between unidimensional and comprehensive nonverbal tests. A unidimensional test is composed of items that all come from a common domain—typically the sort of figural-spatial reasoning required on the Progressive Matrices (Raven et. al., 2000), the NNAT (Naglieri, 1997), and the Nonverbal Battery of CogAT (Lohman & Hagen, 2001a). Alternatively, comprehensive nonverbal tests sample from a broader domain of test content. Examples include the Leiter International Performance scale (Roid & Miller, 1997), the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998), and the primary-level tests of Form 7 of CogAT (Lohman, 2011) shown in Figure 1 (on page xx).

Form 7 of CogAT differs from earlier editions of CogAT, most importantly in ways that make the primary-level tests more accessible to ELL children. These differences are summarized in Figure 6. On previous editions of CogAT, all four subtests on the primary-level verbal and quantitative and batteries required that students listen to a question that was presented orally by the test administrator and then select a picture that best answered that question. (The two subtests on the nonverbal battery did not use items that required language.) On Form 7, however, only one of the nine subtests on the primary battery (i.e., Sentence Completion) requires comprehension of oral language. This test can be administered in English or Spanish or omitted altogether. Figure 1 (on page xx) showed sample items from the three verbal subtests on Form 7 that demonstrate how a pictorial format can be used instead of a text-based format with young children. Eight of the nine CogAT subtests at grades K-2 measure reasoning without the use of verbal prompts or responses: three subtests use figural spatial content (Figure Matrices, Figure Classification, and Paper Folding), three use quantitative content (Number Analogies, Number Series, Number Puzzles), and two use verbal content (Picture Analogies & Picture Classification). These subtests allow one to estimate the child’s ability to reason “nonverbally” in three content domains: spatial, numerical, and verbal. Importantly, ELL, low SES, and minority students perform as well or better on these picture-verbal and picture-quantitative tests than on figural reasoning tests (Lohman & Gambrell, in press). Three content domains on the Form 7 primary-level CogAT tests (picture-verbal, picture-quantitative, and spatial/figural) provide a much broader measure of ability than earlier editions of CogAT (in which only the Figure Matrices and Figure Classification subtests were presented nonverbally) and other nonverbal tests that use only figure matrices (e.g., Raven or NNAT).

#### MISUSES OF GROUP TESTS

The difference between individually administered ability tests and group administered ability tests is like the difference between prescription and over-the-counter drugs. On the one hand, while non-experts are likely to be misled in their beliefs about the efficacy of drugs obtained from either source, there is at least some semblance of control over the claims about and

usage of prescription drugs. On the other hand, claims made by the purveyors of over-the-counter drugs often exaggerate their efficacy, ignore their negative side effects, and sometimes purposely mislead. Likewise, the restraints on the marketing and use of group-administered ability tests that do not require professional certification to administer are few and rarely enforced. The only real restraints are those that are self-imposed by the integrity of the test author and the company that markets the test.

Misuses of scores on group-administered ability tests abound. Some are errors of commission, others of omission. Since such tests enter prominently into decisions about academic talent that educators must make, we discuss several of the most egregious claims made for these tests.

#### ¶ *Does the test purport to be a panacea?*

Most people believe that ability tests measure (or ought to measure) innate ability unfettered by contaminants such as education, culture, and language. Such an ancient belief is more akin to theories held by novices in physics and other scientific domains (Lohman, 2006). Although the vast majority of experts in measurement do not subscribe to this view, the handful that do gather much attention. Unfortunately, most educators have little training in educational and psychological measurement, and thus have little preparation to dispute such claims. In fact, because they have little training in measurement, most find these beliefs congenial with their own beliefs about ability. Some of the more exaggerated claims about ability tests include:

1. that a test is culture fair, culture free, or, more recently, culture neutral;
2. that it will identify equally proportions of minority and non-minority, poor and rich, or ELL and non-ELL children;
3. that it can give tests scores that are sufficiently reliable for gifted identification with approximately 30 minutes of testing; and
4. that pantomime or pictorial directions can adequately prepare students to do their best on these tests.

#### ¶ *Does the test provide instructionally useful information for all students?*

Administrators of talent development programs often encourage their schools to administer an ability test to all children so that talent identification is not contingent on teacher or parent nomination. Census testing also enables the program to obtain local norms. Scores from the screening test serve an obvious purpose for those students who are subsequently offered enriched or more challenging instruction. But what about the other 95% of the children? Does the test give these children, their parents, or their teachers educationally useful information? If not, is it any wonder that teachers might oppose taking precious class time to administer the ability test? Misinterpreting test scores for low-scoring students can be even more problematic if the test is thought to provide a culture-fair measure of innate ability.

#### ¶ *Does the test report errors of measurement that apply to gifted students?*

Some tests report only reliability coefficients and do not present standard errors of measurement (SEM). Test users commonly over-estimate the dependability of scores when given only reliability coefficients (see page xx). Others report SEMs, but not on the score scale that test users will need. For example, InView reports SEMs for number correct scores rather than for the IQ-like scores used for talent identification, which is not helpful. Furthermore, errors of measurement for IQ-like scores are typically many times larger at the extremes of the distribution than at the mean. Thus, the average or typical SEM substantially overestimates the precision of extreme scores, which has obvious implications for using the test to identify gifted students.

### ¶ *Do score reports warn the user when scores for an examinee may be undependable?*

Given the likelihood that scores on ability tests will be misinterpreted—if not by the user then by someone else reading the student’s file—aberrant scores either should not be reported or should be marked as possibly erroneous. Although there are a few cases in which the mistake could result in an over-estimate of ability (e.g., coding a lower age or grade than was appropriate), most testing aberrations reduce test scores. There are many more ways to get something wrong than to get it right. For these reasons examiners or test administrators should report behavioral observations that could negatively influence performance. Although some group administered tests (e.g., NNAT-2 and CogAT) will not report a score if the age score seems unusual, only CogAT cautions scores if the student responded inconsistently to items or subtests that measure the same ability, appeared to have adopted a very slow and cautious response style, or exhibited other obvious problems. Warnings caution users not to make high-stakes decisions about the child using that test score. All computer-scored, group-administered tests should provide these sorts of warnings.

### ¶ *Are the test norms recent?*

The scores on ability tests have risen dramatically over the past 70 years. Performance on IQ tests has been improving ever since ability tests were first introduced. Flynn (2007) estimated an increase of about three IQ points per decade between 1948 and 2002, with the largest increase occurring on figural reasoning tests such as the Progressive Matrices Test (Raven, Court, & Raven, 1996). The mean IQ score remains 100 only because intelligence tests are re-normed every few years; therefore, old norms are invariably too lenient. For example, the 2000 edition of the Progressive Matrices test uses normative data collected in the 1970s and 1980s. A comparison of these normative scores with two more recently normed tests showed that scores on the Raven were approximately 10 IQ-like points (.67 SD) too high (Lohman, Korb, & Lakin,). Normative scores on the Culture-Fair Intelligence test (Cattell & Cattell, 1965), which has not been normed since the 1960s, are even more out of date. One study found that its normative scores were about 17 IQ-like points too lenient (Shaunessy, Karnes, & Cobb, 2004).

### ¶ *Are the norms dependable?*

On other tests, the problem is not the recency of the norms, but their dependability. Norms can be undependable if the norming sample is not representative of the population. Several studies now suggest that the normative sample for the SB-V may have over-represented exceptional children. If this is true, then normative scores on the test would be too high for low-scoring children and too low for high-scoring children. This phenomenon was first suggested in the SB-V technical manual (Roid, 2003). Standard deviations for the SB-V were smaller than standard deviations on the SB-IV, the WIPPSI-R, the WISC-III, the WAIS-III, and the WJIII COG. Although some of the differences were small, others were not. Test users noted a similar restriction in the range of scores on the SB-V, with fewer students obtaining very high scores (Minton & Pratt, 2006). This situation is troubling because, in all other respects, test construction exemplified best practices in psychometrics.

Norms derived retrospectively from user data, rather than purposefully obtained through a sampling plan, are always suspect, even if based on very large samples. Those who choose to purchase and administer a particular test are not a random sample of all potential test users. The normative data for the Progressive Matrices test were obtained in this way. Even when a national sampling plan is used, many of the schools or individuals who are contacted refuse to participate. When test users refuse to participate, replacements are sought or data for other participants who were in the same cell of the sampling design are given greater weight. This tactic can bias the norms that are obtained. Such refusals are increasingly common in an educational system overburdened with testing, even when schools are offered substantial financial incentives for participation. As a result, trustworthy national norms on ability tests are increasingly difficult to obtain.



Norms can also be undependable if the procedures used to develop them were either suboptimal or incorrect. The 1997 norms for the NNAT exhibited both problems. The norming procedures were suboptimal in that the within-age score distributions were not smoothed. Smoothing ensures that IQ-like scores do not change erratically as one moves across age groups. The NNAT norming procedures were incorrect in that the standard deviation of the IQ-like score (called the nonverbal ability index or NAI) was substantially greater than 15 at all but one test level. Inferences about giftedness depend critically on the standard deviation of scores. Excessive variability of NAI scores means that the test over-identified the number of students receiving high scores. For example, the number of students who received NAI scores of 130 or higher on Level A of the NNAT was over three times greater than it should have been. Although one would hope that these problems were corrected in the norming of the NNAT-2, they exemplify the sort of problems that test users should not have to worry about when they purchase a recently-normed test from a major publishing company.

#### ¶ *Does the test provide practice materials?*

Since the early days of testing, psychologists have endeavored to measure intelligence by seeing how well the examinee adapted to the demands of novel problems. To this day, ability tests rely heavily on item types and formats that are unfamiliar to most students. The problem is that not all students are equally unfamiliar with the item formats used on the test. In some cases, unfamiliarity leads to serious misconceptions about how to solve items. In other cases, it leads to sub-optimal performance, especially when the test is speeded. Alternatively, performance is enhanced for those who have practiced using the item formats. How large are these practice and training effects? Practicing unfamiliar item formats commonly results in improvements of 8 or more IQ points. Effects are largest for nonverbal tests and smallest for verbal tests.

There are several ways to reduce these effects. First, those who have custody of tests need to treat the responsibility seriously. Lax security of test booklets increases the likelihood that booklets will be stolen or copied. Copying tests is not only unfair, but illegal. Second, whenever possible, students should be retested with an alternate form of the test. Some tests have alternate forms; most do not. Even when more than one form of the test is available (e.g., Otis-Lennon forms 7 and 8 or NNAT and NNAT-2), tables are not provided that show how scores on the old form map on to the scores on the new form, so that one can use the most recent norms tables for interpreting both sets of scores. Some widely used tests (e.g. the Progressive Matrices Test) have used the same items since they were first developed. Others (e.g., the Otis-Lennon) replace some items and repeat others. When forms use the same items, lax security on one form compromises other forms. Third, students should receive practice in solving items that use the item formats that they will see on the tests. Savvy parents can now buy practice test materials on the internet. The availability of these materials has significantly enhanced the likelihood that children who have access to practice materials will obtain high scores on the tests. Practice for all children is necessary to level the playing field. Good practice should do more than rehearse a few items; it should help students learn how to think about the items, not merely how to respond to them. For example, learning to use language to label stimuli and rules that connect them can be critical for performance on nonverbal tests.

## SUMMARY

Ability testing is a critical component of any academic talent identification system. It is most important for those who, though age, experience, circumstance, or choice have not developed high levels of competence in some academic or cognitive domain. This is commonly true for young children, for poor and minority students whose circumstances have limited their

opportunities to acquire academic skills, and for twice-exceptional children, whose disability may have negatively affected their classroom performance. Although ability data is critical, it should never be expected to stand alone. Furthermore, the abilities that should be measured cannot adequately be assessed in a few minutes by administering two or three subtests from a larger assessment battery, or worse, a single kind of test-task. Decision makers who use brief, unidimensional tests are often misled by seemingly high reliabilities. High internal consistency reliabilities mean only that the items on the test measure the same thing, not that they generalize to anything else. Tests need multiple formats and contents for greater generalizability. Further, instead of reliability coefficients, information about standard errors of measurement and how they vary depending on ability level are more useful.

Test interpreters must move beyond overall composite scores and other measures of  $g$  in order to understanding how well students reason in different symbol systems commonly required for academic learning: verbal, spatial, and quantitative. This is particularly true for high ability students who are more likely to have uneven ability profiles than students who score closer to the mean. When selecting ability tests that will most fairly and accurately assess student's abilities, it is important to go beyond simple differences in the mean scores between groups and attend to issues of predictive validity, or the extent to which the test measures those aptitudes that are necessary for and thus predictive of success in the talent domain. In other words, the assessment tools must measure skills that will be needed in the talent development program. Although concerns about equity can in part be addressed by using comprehensive nonverbal reasoning tests, no test can measure innate ability in a way that is independent of education, culture, and experience. Developing talent identification systems that allow ability test scores to be interpreted using different norm groups -- national, local, and opportunity-to-learn -- usually offers a better way to achieve equity without compromising excellence.

There are individual as well as group administered ability tests, and each has its own set of advantages and disadvantages. Individually administered tests have the advantage of allowing for observations of behaviors that influence scores, generally well-developed normative data, better understanding of twice-exceptional students, and well-trained examiners to interpret test data. At the same time, the use of individually-administered tests as a part of an identification program for gifted students raises questions about equity, cost-effectiveness, and efficiency. Group ability tests address these limitations because they can be administered quickly to large groups of students, and local norms can be easily developed for comparisons that better account for differences in opportunity to learn. However, many group-administered tests are poorly developed, have inadequate norms, or are advertised in misleading ways. Frank discussions with current users of a test who have looked carefully at their test data and how well it has lived up to expectations can be helpful in selecting a test to administer.

Misconceptions about ability abound. Parents, teachers, and even professionals not explicitly trained in assessment have naïve beliefs about what these tests measure, which makes it imperative that those who administer ability tests not use them in ways that reinforce these misconceptions.

## REFERENCES

- Ackerman, P. L. (2003). Aptitude complexes and trait complexes. *Educational Psychologist, 38*, 85-94.
- Achter, J. A, Benbow, C. P., & Lubinski, D. (1997). Rethinking multipotentiality among the intellectually gifted: A critical review and recommendations. *Gifted Child Quarterly, 41*, 5-15.
- Anastasi, A. (1937). *Differential psychology*. New York: Macmillan.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Assouline, S. G. (2003). Psychological and educational assessment of gifted children. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 124-145). Boston: Allyn & Bacon.
- Assouline, S. G., Colangelo, N., Lupkowski-Shopluk, A., Lipscomb, J., & Forstadt, L. (2009). *Iowa acceleration scale*, third edition. Scottsdale, AZ: Great Potential Press.
- Assouline, S. G., Foley Nicpon, M., & Bramer, D. M. (2006). The impact of vulnerabilities and strengths on the academic experiences of twice-exceptional students: A message to school counselors. *Professional School Counseling, 10* (1), 14-25.
- Assouline, S. G., Foley Nicpon, M., & Whiteman, C. S. (2010). Cognitive and psychosocial characteristics of gifted students with written language disability. *Gifted Child Quarterly, 54*, 102-115. Doi: 10.1177/0016986209355974. Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York: Harper & Brother.
- Borland, J. H. (2004). Issues and practices in the identification and education of gifted students from under-represented groups RM04186). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.
- Bracken, B. A., & McCallum, S. (1998). *Universal Nonverbal Intelligence Test - UNIT*. Itasca, IL: Riverside Publishing.
- Braden, J. P. (2000). Perspectives on the nonverbal assessment of intelligence. *Journal of Psychoeducational Assessment, 18*, 204-210.
- Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Mind, brain, experience, and school*. Washington, DC: National Academy Press.
- Carmen, C. A, & Taylor, D. K. (2010). Socioeconomic status effects on using the Naglieri Nonverbal Ability Test (NNAT) to identify the gifted/talented. *Gifted Child Quarterly, 54*, 75-84.
- Carroll, J. B. (1974). The aptitude-achievement distinction: The case of foreign language aptitude and proficiency. In D. R. Green (Ed.), *The aptitude-achievement distinction* (pp. 286-303). Monterey, CA: CTB/McGraw-Hill.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.
- Cattell, R. B., & Cattell, K. S. (1965). *Manual for the Culture-Fair Intelligence Test, Scale 2*. Champaign, IL: Institute for Personality and Ability Testing.
- 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. & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456-473.
- Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.
- Daniel, M. H. (2000). Interpretation of intelligence test scores. In R. Sternberg (Ed.), *Handbook of intelligence* (pp. 477-491). New York: Cambridge.
- Feldhusen, J. F. & Jarwan, F. A. (2000). Identification of gifted and talented youth for educational programs. In K. A. Heller, F. J. Monks, R. Subotnik, & R. J. Sternberg (Eds.) *International handbook of giftedness and talent* (pp. 271-282). Oxford,

UK: Elsevier

- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.
- Flynn, J. (2007). *What is intelligence: Beyond the Flynn effect*. NY: Cambridge University Press.
- 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*, 35(2), 169-182.
- Gagné, F. (2009). Talent development as seen through the differentiated model of giftedness and talent. The Routledge international companion to gifted education. In T. Balchin, B. Hymer, & D. J. Matthews (Eds). *The Routledge international companion to gifted education*. (pp. 32-41). New York, NY, US: Routledge/Taylor & Francis Group.
- 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.
- Gohm, C. L., Humphreys, L. G., & Yao, G. (1998). Underachievement among spatially gifted students. *American Educational Research Journal*, 35, 515-531.
- Gudjonsson, G. H. (1995). The Standard Progressive Matrices: Methodological problems associated with the administration of the 1992 adult standardization sample. *Personality and Individual Differences*, 18, 441-442.
- Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, 28, 407-434.
- Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms*. New York: Cambridge University Press.
- Hernstein, R. J. & Murray, C. (1994). *The bell curve: intelligence and class structure in American life*. New York : Free Press.
- Hoover, H. D., Dunbar, S. B., & Frisbie, D. A., (2005). Iowa Tests of Basic Skills (Forms A, B, and C). Itasca, IL: Riverside.
- Hunt, E. (2000). Let's hear it for crystallized intelligence. *Learning and Individual Differences*, 12, 123-130.
- Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. Wiley New York.
- Kaufman, A. S., & Harrison, P. L., (1986). Intelligence tests and gifted assessment: What are the positives? *Roeper Review*, 8(3), 154-159.
- Kaufman, S. B., & Sternberg, R. J. (2008). Conceptions of giftedness. In S. I. Pfeiffer (Ed.), *Handbook of giftedness in children: Psychoeducational theory, research, and best practices* (pp. 71-91). New York, NY: Springer Science + Business Media.
- Kaufman, A. S., & Sternberg, R. J. (2007). Giftedness in Euro-American culture. In S. N. Phillipson & M. McCann (Eds.), *Conceptions of giftedness: Socio-cultural perspectives*. Mahwah, NJ: Erlbaum.
- 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. (1994). Spatial ability. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 1000-1007). New York: Macmillan.
- 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](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%20r.pdf](http://faculty.education.uiowa.edu/dlohman/pdf/CogAT_WJIII_final_2col%20r.pdf)
- Lohman, D. F. (2005a). An aptitude perspective on talent identification: Implications for the identification of academically gifted minority students. *Journal for the Education of the Gifted*, 28, 333-359.
- 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. (2006). Beliefs about differences between ability and accomplishment: From folk theories to cognitive science. *Roeper Review*, 29, 32-40.
- Lohman, D. F. (2011). *Cognitive Abilities Test (Form 7)*. Rolling Meadows, IL: Riverside.
- Lohman, D. F., & Hagen, E. P. (2001a). *Cognitive Abilities Test (Form 6)*. Itasca, IL: Riverside.
- Lohman, D. F., & Hagen, E. P. (2001b). *Cognitive Abilities Test (Form 6): Interpretive guide for teachers and counselors*. 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. (2008). Identifying academically gifted English language learners using nonverbal tests: A comparison of the Raven, NNAT, and CogAT. *Gifted Child Quarterly*, 52, 275-296.
- Lohman, D. F., & Gambrell, J. L. (in press). Use of nonverbal tests in gifted identification. *Journal of Psychoeducational Assessment*.
- Lohman, D. F., Gambrell, J., & Lakin, J. (2008). The commonality of extreme discrepancies in the ability profiles of academically gifted students. *Psychology Science Quarterly*, 50, 269-282.
- Lohman, D. F., & Lakin, J. (2007). Nonverbal test scores as one component of an identification system: Integrating ability, achievement, and teacher ratings. In J. VanTassel-Baska (Ed.), *Alternative assessments for identifying gifted and talented students* (p. 41-66). Austin, TX: Prufrock Press.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N.J.: Erlbaum Associates.
- Lubinski, D. (2004). Introduction to the Special Section on Cognitive Abilities: 100 years after Spearman's (1904) "General Intelligence, Objectively Determined and Measured." *Journal of Personality and Social Psychology*, 86(1) 96-111.
- Lubinski, D., & Benbow, C. P. (2000). States of excellence. *The American Psychologist*, 55(1), 137-150.
- Lubinski, D., Webb, R. M., Morelock, M. J., & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology*, 86(4), 718-129.
- McCallum, S., Bracken, B., & Wasserman, J. (2001). *Essentials of Nonverbal Assessment*. Hoboken, NY: Wiley.
- Mills, C., & Tissot, S. L. (1995). Identifying academic potential in students from under-represented populations: Is using the Ravens Progressive Matrices a good idea? *Gifted Child Quarterly*, 39, 209-217.
- Minton, B. A., & Pratt, S. (2006). Gifted and highly gifted students: How to they score on the SB5? *Roeper Review*, 28, 232-236.
- Mulaik, S. A. (1972). *The foundations of factor analysis*. New York, NY: McGraw-Hill.
- National Association for Gifted Children (2008). Use of the WISC-IV for gifted identification. Retrieved from [http://www.nagc.org/uploadedFiles/Information\\_and\\_Resources/Position\\_Papers/WISC-IV.pdf](http://www.nagc.org/uploadedFiles/Information_and_Resources/Position_Papers/WISC-IV.pdf)
- Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test: Multilevel technical manual*. San Antonio, TX: Harcourt Brace.
- Naglieri, J. A., & Ford, D. Y. (2003). Addressing underrepresentation of gifted minority children using the Naglieri Nonverbal Ability Test (NNAT). *Gifted Child Quarterly*, 47, 155-160.
- Naglieri, J. A., & Ronning, M. E. (2000). The relationship between general ability using the Naglieri Nonverbal Ability Test (NNAT) and the Stanford Achievement Test (SAT) reading achievement. *Journal of Psychoeducational Assessment*, 18, 230-239.
- Naglieri, J. A., Booth, A. L., & Winsler, A. (2004). Comparison of Hispanic children with and without limited English

- proficiency on the Naglieri Nonverbal Ability Test. *Psychological Assessment*, 16, 81-84.
- Newman, T. M., Sparrow, S. S., & Pfeiffer, S. I. (2008). The use of the WISC-IV in assessment and intervention planning for children who are gifted. In A. Prifitera, D. H. Saklofske, & L. G., Weiss (Eds.), *WISC-IV clinical assessment and intervention*. San Diego, CA: Academic Press.
- Park, G., Lubinski, D., & Benbow, C. P. (2007). Contrasting intellectual patterns predict creativity in the arts and sciences: Tracking intellectually precocious youth over 25 years. *Psychological Science*, 18, 948-952.
- Petersen, P. (1977). Interactive effects of student anxiety, achievement orientation, and teacher behavior on student achievement and attitude. *Journal of Educational Psychology*, 69, 779-792.
- Powers, S., Barkan, J. H., & Jones, P. B. (1986). Reliability of the Standard Progressive Matrices Test for Hispanic and White-American children. *Perceptual and Motor Skills*, 62, 348-350.
- Raven, J. 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, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 1. General overview*. Oxford, UK: Oxford Psychologists Press Ltd.
- 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.
- Rimm, S., Gilman, B., & Silverman, L. (2008). Nontraditional applications of traditional testing. In J. L. VanTassel-Baska (Ed.), *Alternative assessments with gifted and talented students*. Wako, TX: Prufrock Press.
- Robinson, N. M. (2008). The value of traditional assessments as approaches to identifying academically gifted students. In J. L. VanTassel-Baska (Ed.), *Alternative assessments with gifted and talented students*. Wako, TX: Prufrock Press.
- 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.
- Roid, G. (2003). *Stanford Binet Intelligence Scales, technical manual (5th ed.)*. Itasca, IL: Riverside.
- Roid, G. H., & Miller, L. J. (1997). *Leiter International Performance Scale-Revised*. Wood Dale, IL: Stoelting Co.
- Saccuzzo, D. P., Johnson, N. E., & Russell, G. (1992). Verbal versus performance IQs for gifted African-American, Caucasian, Filipino, & Hispanic children. *Psychological Assessment*, 4, 239-244.
- Scarr, S. (1994). Culture-fair and culture-free tests. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 322-328). New York: Macmillan.
- Shaunessy, E., Karnes, F. A., & Cobb, Y. (2004). Assessing potentially gifted students from lower socioeconomic status with nonverbal measures of intelligence. *Perceptual and Motor Skills*, 98, 1129-1138.
- Schoen, H. L., & Ansley, T. N. (2005). *Iowa Algebra Aptitude Test, (5th edition)*. Itasca, IL: Riverside.
- Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology*, 93, 604-614.
- Silverman, L. K. (2009). The measurement of giftedness. In L. V. Shavinina (Ed.), *International handbook on giftedness*. New York, NY: Springer Science + Business Media.
- Snow, R. E., & Lohman, D. F. (1984). Toward a theory of aptitude for learning from instruction. *Journal of Educational Psychology*, 76, 347-376.
- Snow, R. E., & Yalow, E. (1982). Education and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp.

- 493-585). Cambridge, England: Cambridge University Press.
- Stanley, J. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd Ed.). Washington, DC: American Council on Education.
- Sternberg, R. J. (2003). Giftedness according to the theory of successful intelligence. N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed, pp. 88-99). Boston: Allyn & Bacon.
- Subotnik, R., & Jarvin, L. (2005). Beyond expertise: Conceptions of giftedness as great performance. In R. J. Sternberg & J. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 343-57). New York: Cambridge University Press.
- Sweetland, J. D., Reina, J. M., & Tatti, A. F. (2006). WISC-III verbal/performance discrepancies among a sample of gifted children. *Gifted Child Quarterly*, 50(1), 7 – 10.
- Terman, L. M. (1930). Autobiography of Lewis M. Terman. In Murchison, C. (Ed.) *History of psychology in autobiography* (Vol. 2, pp. 297-331). Worcester, MA: Clark University Press.
- Thorndike, R. L. (1963). *The concepts of over- and under-achievement*. New York, Bureau of Publications, Teachers College, Columbia University.
- Thorndike, R. L., & Hagen, E. (1971). *Cognitive Abilities Test*. New York: Houghton Mifflin.
- Thorndike, R. L., & Hagen, E. (1978). *Cognitive Abilities Test (Form 3)*. New York: Houghton Mifflin.
- Thorndike, R. L., & Hagen, E. (1986). *Cognitive Abilities Test (Form 4): Examiner's manual*. Chicago: Riverside.
- Thorndike & Hagen, 1987). *Cognitive Abilities Test (Form 4): Research handbook* .Chicago: Riverside.
- Thorndike, R. L., & Hagen, E. (1994). *Cognitive Abilities Test (Form 5): Interpretive guide for teachers and counselors*. Chicago: Riverside
- Thorndike, R. L., & Hagen, E. (1995). *Cognitive Abilities Test (Form 5): Research handbook*. Chicago: Riverside.
- VanTassel-Baska, J., & Brown, E. F. (2007). Toward best practice: An analysis of the efficacy of curriculum models in gifted education. *Gifted Child Quarterly*, 51(4), 342-358.
- Watkins, M.W., Gluting, J.J., & Youngstrom, E. A. (2005) Issues in subtest profile analysis. In D. P. Flanagan & P. L. Harrison (Eds.) *Contemporary intellectual assessment* (2nd ed., pp. 251-268). New York: The Guilford Press
- Wechsler WISC IV Spanish
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: The Psychological Corporation.
- Williams, P. E., Weiss, L. G., & Rolhus, E. L. (2003). *Wechsler Intelligence Scale for Children – Fourth Edition Clinical Validity. Technical Report #3*. San Antonio, TX: Pearson Education Inc.
- Wilkinson, S. C. (1993). WISC-R profiles of children with superior intellectual ability. *Gifted Child Quarterly*, 37, 84-91.
- Zhu, Clayton, Weiss, & Gabel, (2008). WISC-IV Extended norms. Technical Report No. 7. Pearson Education, Inc.

Table 1

Standard Errors of Measurement for a 10-year old child scoring near the mean on several ability tests

	<b>WISC- IV<sup>a</sup></b>	<b>SB-V</b>	<b>OLSAT -8</b>	<b>Inview<sup>c</sup></b>	<b>Raven<sup>d</sup> SPM</b>	<b>NNAT</b>	<b>CogAT 6</b>
Verbal	3.9	3.6	5.7	5.3			3.4
Nonverbal/Perceptual	4.2	3.9	5.8 <sup>b</sup>	4.5 <sup>b</sup>	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;

<sup>a</sup> Working Memory Composite used to estimate Quantitative for WISC IV

<sup>b</sup> 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

<sup>c</sup> 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.

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

\*\*\*\*\*in PDF version the superscripts are not present and need to be for accuracy\*\*\*\*\*



Box 1

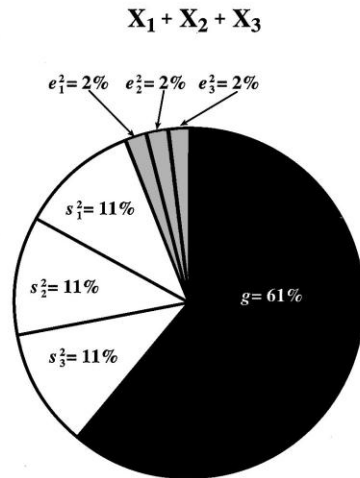
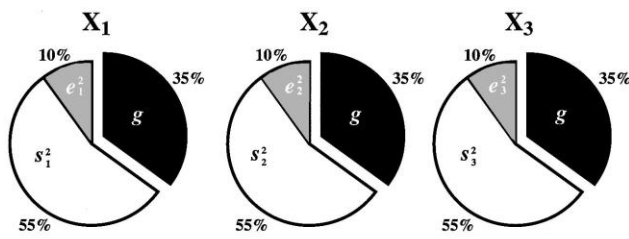
A talent classification scheme for schools

1. *Talents that schools can (or should) develop.* This category includes academic subjects that, by custom or mandate, schools already strive to develop. All elementary schools have well-developed educational programs in literature, writing, mathematics, social studies, and science; most also have at least rudimentary programs in music and the arts. In many of these domains, the level of instruction needed for academically or artistically precocious children may only be offered at more advanced grades. In other cases, the school may not have programs in the domain (e.g., art or music) but could well be lobbied to develop such programs. Access to such programs is thus a different issue than whether schools aim to develop a particular kind of talent.
2. *Talents that schools cannot themselves develop but should encourage through involvement with external organizations.* Teachers observe a wide range of other talents that children and their parents can be encouraged to develop with the assistance of organizations outside of the school. Musical and athletic talents can sometimes be developed in school-based programs. But schools may not have the resources to provide more than an elementary development. Many of the summer and extracurricular programs offered by talent search programs such as John Hopkins University, Northwestern, and the Belin-Blank Center at The University of Iowa provide such enrichment opportunities for academically precocious youth. Increasingly, administrators of these programs have endeavored to make them available to all children, not just those children whose parents can afford them.
3. *Talents that are outside of the school's purview.* Everything from sports clubs to political action groups fall in this category. Talent development in these specific domains certainly should be encouraged when a student exhibits motivation and interest, but they are not directly supported by school resources. In these cases, school support personnel, such as counselors, can assist students with discovering extracurricular activities that can develop and foster their unique interests and talents.

Box 2

The Effects of Averaging (from Lubinski, 2004)

The figure below illustrates what happens when a test with three score scales that are moderately correlated (say quantitative, spatial, verbal) are summed (or averaged). All three test scales ( $X_1$ ,  $X_2$ , and  $X_3$ ), have .90 reliabilities (or 10% random error). For each, the preponderance of their variance measures a specific construct (55%), namely, quantitative, spatial, or verbal ability, but each also has an appreciable general factor component (35%). Aggregation of these three scales results in a composite score that primarily reflects the general factor running through all three indicators (61%). The remaining components of unique variance associated with each indicator shrink to tiny slivers of content homogeneity (11% each) and random error (2% each).



Box 3

Widely Used Individually-administered ability tests

Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV)

- Ages 6.0 – 17.11
- Contains 10 core and 5 supplemental subtests. Core subtests are summed to a full scale IQ and four indices: Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and Processing Speed Index (PSI).
- General-Ability Index (GAI) recommended in many assessment situations, such as when a significant and unusual (base rate of less than 10 - 15%) discrepancy exists between the VCI and WMI, the PRI and PSI, and/or WMI and PSI; or when unusual scatter exists among WMI and/or PSI subtests.
- Extended norms to 210 for composites, and 28 points for subtests; new and requires additional clinical validation; however, extended norms rarely are used.

Stanford-Binet Intelligence Scale, 5<sup>th</sup> Edition (SB-5)

- Ages 2 – 85+
- Contains 10 subtests which are combined into a full scale IQ, two domain scores (Verbal IQ and Nonverbal IQ), and five indices (Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual-Spatial Processing, and Working Memory)
- An experimental Gifted index has been proposed that sums 3 nonverbal and 4 verbal tests. This excludes the NV Visual-Spatial Processing and Working Memory subtests.
- Extended norms to 160 for composites; experimental and supplemental; however, extended norms rarely are used. dcock-Johnson III Tests of Cognitive Abilities (WJ III COG)
- Ages 2 – 90+
- Contains 10 standard and 10 supplemental tests. These are summed to give multiple indices: General Intellectual Ability (based on 7 standard tests), General Intellectual Ability (based on 7 standard and 7 supplemental tests), and Brief intellectual Ability (based on 3 standard tests). Other scores are reported for Cognitive Categories (6 indices), CHC Factors (7 indices), and Clinical Clusters (7 indices).
- Range of Standard Scores for total test composite: 0 – 200.

Box 4								
Commonly Used Group Ability Tests								
	Scores	Grades	Items per Battery	Total Items	Testing Time (min)	Score Warnings	Conditional Errors of Measurement	Most recent norms
Cognitive Abilities Test (CogAT) (Form 7)	Verbal, Quantitative, & Nonverbal	K - 12	38 to 62	118 to 176	90	9	yes - for scale scores	2010
INVIEW	Verbal, Quantitative, & Nonverbal	2 - 12	20 to 40	100	95	0	yes- for raw scores	2000
Naglieri Nonverbal Ability Test (NNAT-2)	Nonverbal		30	30	30	0	no	2007
Otis-Lennon School Ability Test (OLSAT) (Form 8)	Verbal & Nonverbal	K - 12	30 to 36	60 to 72	60	0	no	2002
Standard Progressive Matrices (Raven)	Nonverbal	3 - 12	60	60	untimed	0	no	1970's user norms

Box 5

Estimating Expected Regression in Test Scores

Equation 1 can be used to estimate the expected regression in status scores such as IQs. Note that regression effects for gifted students can be offset by practice effects, which can be substantial for tests that use novel formats or are speeded. Studies that report no regression in the average score for a group usually do not account for general improvements due to practice.

The first step is to convert the IQ to a z score by subtracting the mean IQ and dividing by the population *SD* for the test. For example, if the mean is 100 and the *SD* is 15, then an IQ of 130 converts to a z score of  $= 2.0$ .

If the correlation between scores at time 1 and time 2 is  $r = .8$ , then the expected z score at time 2 is  $2.0 \times .8 = 1.6$ .

This converts to an IQ of  $(1.6 \times 15) + 100 = 124$ . The expected regression is 6 IQ points.

If the IQ were 145, then the expected regression would be 9 IQ points.

Box 6

Nonverbal Tests

“When general intelligence is the targeted construct, the heavy verbal-demands of most language-loaded tests can create unfair construct-irrelevant influences on the examinees’ performance.” (McCallum, Bracken, & Wasserman, 2001, p. 4)

“Tests (such as the Progressive Matrices or Naglieri Nonverbal Ability test) should not be used interchangeably with traditional intelligence tests in situations in which decisions about eligibility are made.” (McCallum, Bracken, & Wasserman, 2001, p. 9.)

“Non-verbal tests are often misleadingly described as tests of intelligence when, in fact, they sample only certain aspects of intellectual functioning. “ (Raven, Court, & Raven, 1998, p. G70).

“For ...items... such as those on the *Raven Progressive Matrices Test*, understanding the task is more than half the battle. In one study, many of the ethnic minority children in the sample did not understand the instructions to the ‘game’ and thus could not solve the problem.” (Scarr, 1994, p.XX)

“... a growing body of evidence suggests that nonlanguage tests may be more culturally loaded than language tests.” (Anastasi & Urbina, 1997, p. 343).

“There is an aspect of problem solving that is clearly rooted in culture, namely the habit of translating pictorial events into sentences and talking about them. Although children may all recognize ovals, triangles, and trapezoids, and may all know about making things bigger or shading them with horizontal rather than vertical lines, the habit of labeling and talking aloud about such things varies across cultures (Heath, 1983). Children who do not actively label objects and transformations are more likely to resort to a purely perceptual strategy on nonverbal tests. Such strategies often work well on the easiest items that require the completion of a visual pattern or a perceptually salient series, but fail miserably on more difficult items that require the identification and application of multiple transformations on multiple stimuli.” (Lohman, 2005b, p.115).

Figure 1. Reasoning subtests on Form 7 of the Cognitive Abilities Test (Lohman, 2011) showing examples of item formats for grades k – 2 (Col 1) and grades 3 – 12 (Col 2.)











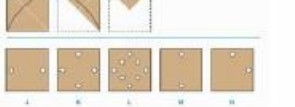


		Picture Format (Grades K – 3)	Text Format (Grades 3 – 12)
<b>VERBAL BATTERY</b>	Verbal Analogies		<p>foot → shoe : hand →</p> <p>mirror   hammer   glove</p> <p><input type="radio"/>   <input type="radio"/>   <input checked="" type="radio"/></p>
	Sentence Completion	<p>"Which one swims in the ocean?"</p> 	<p>A _____ swims in the ocean.</p> <p>cat   shark   bird</p> <p><input type="radio"/>   <input checked="" type="radio"/>   <input type="radio"/></p>
	Verbal Classification		<p>basketball   soccer   football</p> <p>baseball   globe   hoop</p> <p><input checked="" type="radio"/>   <input type="radio"/>   <input type="radio"/></p>
<b>QUANTITATIVE BATTERY</b>	Number Analogies		<p>[1 → 2]   [3 → 4]   [5 → ?]</p> <p>A 2   B 4   C 6   D 8   E 12</p>
	Number Puzzles		<p></p> <p>J 3   K 4   L 5   M 6   N 7</p>
	Number Series		<p>1 2 4 5 7 8 →</p> <p>A 7   B 6   C 9   D 10   E 11</p>
<b>NONVERBAL BATTERY</b>	Figure Matrices		
	Paper Folding		
	Figure Classification		

Figure 2. Profile frequency by median stanine. Flat or “A” profiles (diamonds); a significant strength (B+) or extreme strength [E(B+)] (squares); a significant weakness (B-) or extreme weakness [E(B)] (triangles)

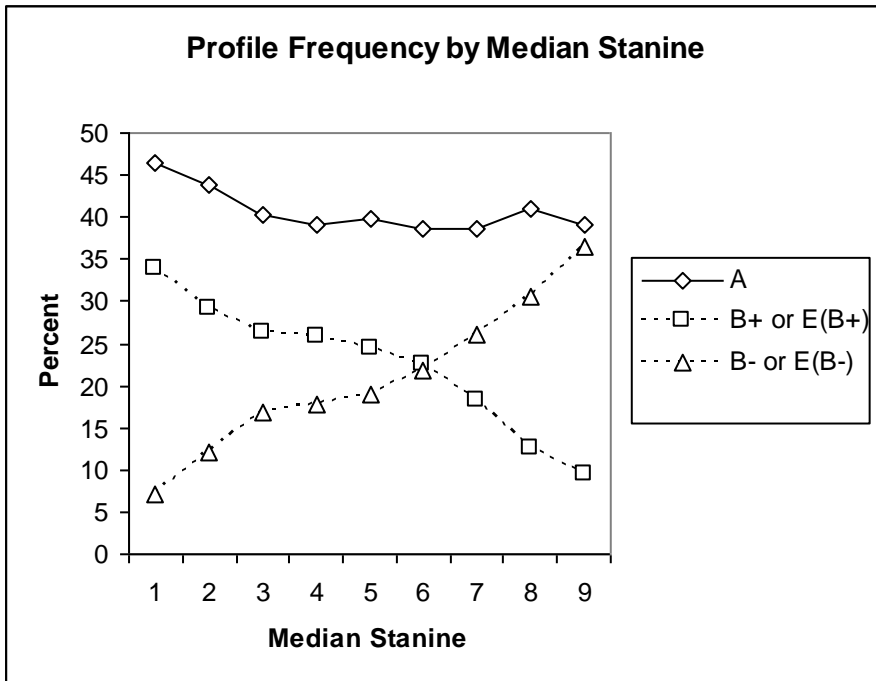
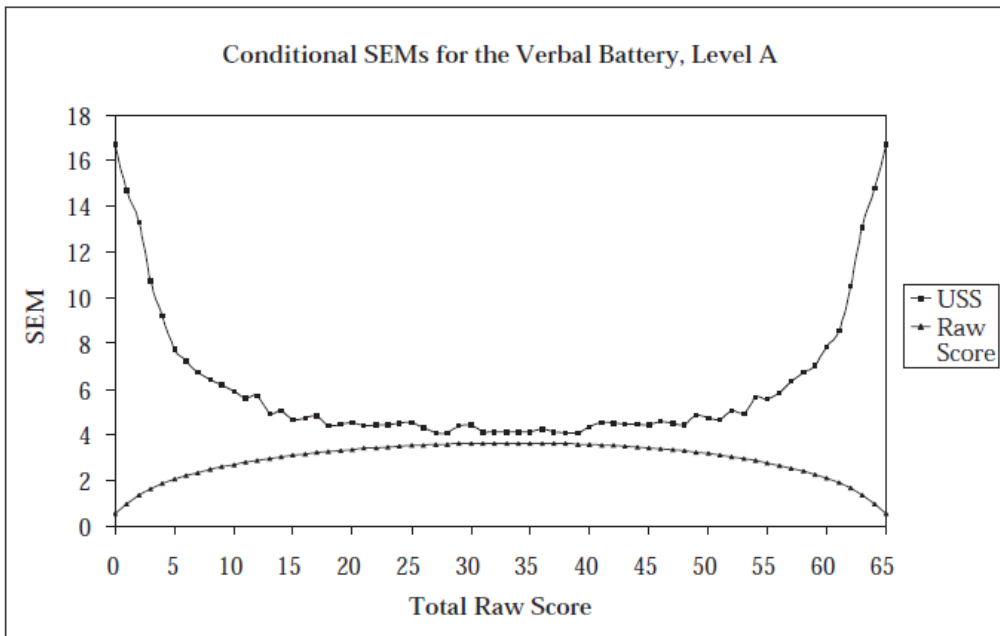




Figure 3. A comparison of errors of measurement for total raw scores (number correct) and the corresponding Universal Scale Scores (USS) for the Verbal Reasoning Battery of CogAT-Form 6 (from Lohman & Hagen, 2002, p. 58).



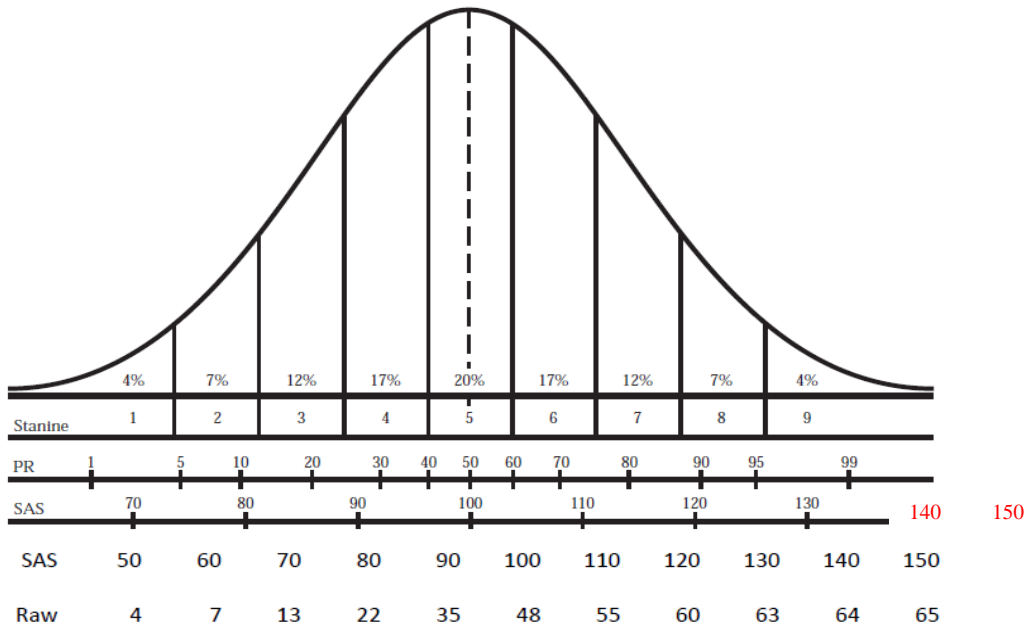
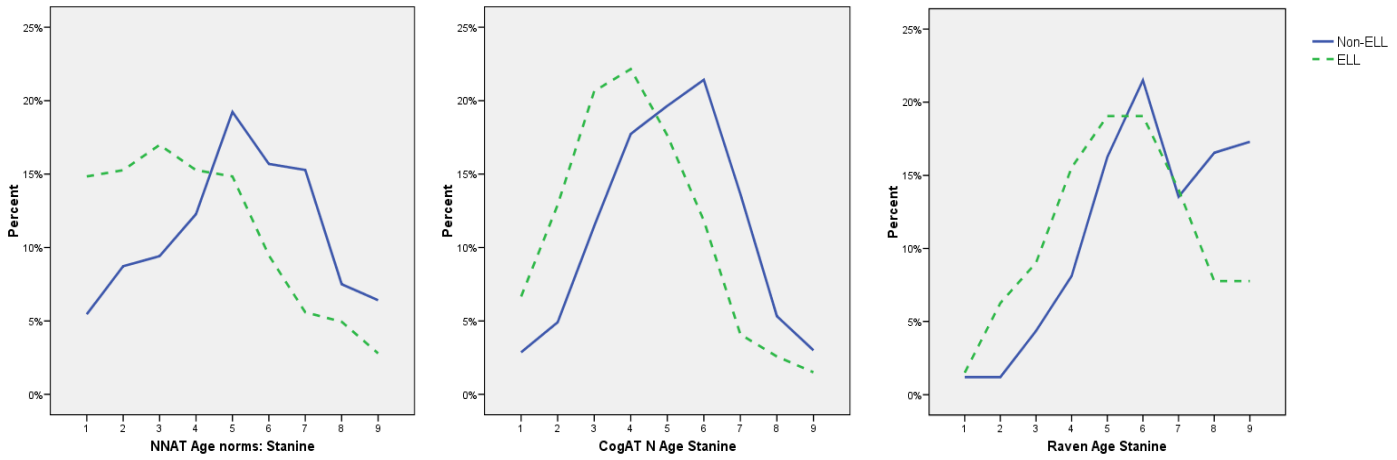


Figure 4. Relationships among Percentile ranks (PRs), Standard Age Scores (SAS), and raw scores for Level B of the CogAT (Form 6) Verbal Battery.



*Figure 5.* Percent of ELL and non-ELL students at each stanine on the Naglieri Nonverbal Ability Test (NNAT; left panel), the Cognitive Abilities Test Nonverbal Battery (CogAT N; center panel), and the Standard Progressive Matrices (Raven; right panel). The dashed green line is for ELL students and the solid blue line for non-ELL students. Note the large number of low-scoring ELL children in Panel 1 and the large number of high-scoring non-ELL children in Panel 3. Both lines should approximate a normal or bell-shaped curve. Curves would overlap if scores for ELL students did not differ from the scores of non-ELL students.

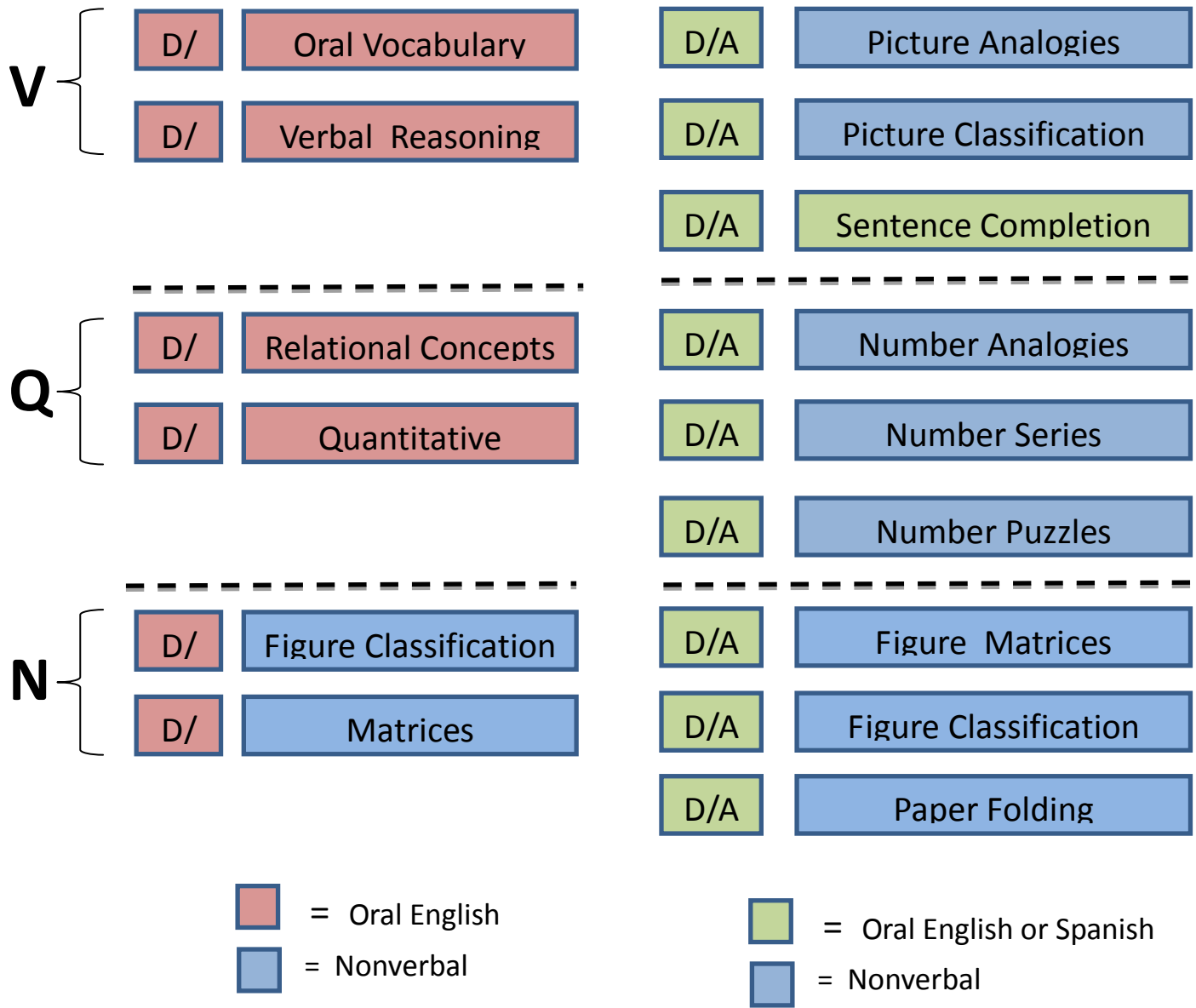


Figure 6. Comparison of subtests on the Primary-level tests (Grades K, 1, and 2) of Forms 6 and 7 of CogAT. Both have Verbal (V), Quantitative (Q), and Nonverbal (N) batteries, each of which has two subtests on Form 6 and three subtests on Form 7. Total testing time is slightly less for Form 7. Both English and Spanish Directions for Administration (D/A) fare provided for the paper-and-pencil version of CogAT7. Several other widely used languages are also provided as audio files on the computer-administered version of Form 7.