Nonverbal test scores as one component of an identification system: Integrating ability, achievement, and teacher ratings.

David F. Lohman

Joni Lakin

University of Iowa

August 23, 2006

There is always an easy solution to every human problem--neat, plausible, and wrong.

Mencken, H. L. (1949, p. 443).

This paper is based in part on an invited presentation by the first author at the Eight Biennial Henry B. & Jocelyn Wallace National Research Symposium on Talent Development, May 2006, University of Iowa, Iowa City

Poor and minority students are underrepresented in programs that serve academically advanced and talented students. Because of this, program administrators have searched for alternative procedures for identifying academic talent that would increase the ethnic, linguistic, and socio-economic diversity of children in their programs. Nonverbal ability tests can help educators achieve these goals. Indeed, the best individual and group-administered ability tests have long incorporated nonverbal subtests (e.g., Lorge & Thorndike, 1954; Wechsler, 1949). However, the nonverbal score has always been viewed as a supplementary measure of ability for examinees who are native speakers of the language in which the test is administered and as a surrogate measure of ability for those who are deaf or who are not native speakers of the language. Nonverbal tests are not simply verbal tests shorn of language. Rather, the constructs that they can measure necessarily under-represent the ability constructs measured by similar tests that use words. This is because eliminating language also eliminates much of what we mean by intelligent thinking.

In this chapter, we describe an approach for identifying academically talented students from all backgrounds that incorporates nonverbal tests as one part of a more comprehensive talent identification system. Five principles guide the approach. First, to identify the right students one must measure the right aptitudes. The development of different kinds of expertise requires different aptitudes. Therefore, one must specify the kind of competence that one hopes students will develop and the demands that educational systems which develop that competence place on them. Second, different inferences from test scores require different comparison or norm groups. Common norms and standards are appropriate for inferences about a student's level of development in a domain. However, inferences about aptitude require comparisons to others who have had similar opportunities to acquire the knowledge and skills measured by a test. Using common norms to estimate the abilities of all students – regardless of their opportunities to learn – leads either to the use of tests that are inferior measures of academic aptitude or to the identification of very few minority students. Third, the best procedures for identifying academically talented students combine different sources of information – such as ability test scores, achievement test scores, and teacher ratings – in a principled way that is guided by research. Fourth, students of the same age who are inferred to have particular academic talents often have markedly different instructional needs. An undifferentiated label such as "gifted" does not usefully guide decisions about the kind of instruction students need, especially as they mature. Fifth, the label "gifted" implies a permanent superiority that misleads. The majority of children who obtain high scores on ability or achievement tests do not retain their status for more than a year or two. Each year, new children excel. Others whose accomplishments were unusual at one age show less precocity a few years later. Therefore, identification of unusual talent and accomplishment should be an ongoing activity.

Measure the Right Aptitudes

For many years the late Richard Snow tried to convince educational researchers of the importance of the concept of aptitude.[1] He defined the concept of aptitude much more broadly than most people would define it. As he used the term, *aptitude* implies a readiness to learn or to perform well in a particular situation. This means that the person is not only capable of performing well in the situation, but is actually in tune with it. There is a helpful reciprocity between what the situation demands or makes possible and what the person brings to it. The attainment of a high level of competence in any domain requires many different kinds of personal resources – some cognitive (e.g., reasoning abilities, prior knowledge and skill), some affective (e.g., motivation, interest), and some conative (e.g., persistence). The particular mix of aptitudes required for success varies across disciplines, and, within a discipline, as the learner gains competence in it. Indeed, one of the most important features of an aptitude perspective is that it goes beyond simplistic talent identification systems that ignore interest, motivation, perseverance, anxiety, or even accumulated knowledge and skill in a domain. *Aptitude* is thus a word that is similar to – but much broader than – the word *talent*. As such, it meshes better with programs that aim to identify and develop specific talents rather than those programs that first identify those who are “gifted” and only secondarily seek to discover what those gifts might be.

Aptitude cannot be understood apart from either the kind of learning that must occur or the contexts in which it must take place. An aptitude perspective begins not with the person but with the kind of expertise that is to be developed. Is the goal to become a writer? A research chemist? A mathematician? Each requires not only the ability to learn different kinds of knowledge and skills, but interest and the ability to persist in the pursuit of excellence. Next, one must understand the demands and affordances of educational systems that students must negotiate if they are to develop this competence. Thriving in classes that require much independent learning requires different personal resources than thriving in more structured classes. Being the youngest person, the only female, or the only Black student in the class requires other personal resources. Changing the demands of the learning situation (e.g., from discovery to didactic teaching) also changes the likelihood that a student will succeed. Aptitudes are thus not free-floating; rather, they are tied to situations.

Most talent identification systems are far more restricted in the model of readiness that guides their identification procedures and therefore in the kind of information that they collect. Some collect much information on each student, but have no empirically substantiated way of combining it to identify those students who are most likely someday to attain expertise. This turns out to be critical for the identification of academically talented minority students. Put differently, different procedures will be used if the goal is to admit more minority students to a program than if the goal is to identify those minority students who are most likely to succeed in that program.

It is sometimes asserted that giftedness manifests itself in different ways in minority students than in non-minority students. Although there is surely true to some degree, on the whole there is much more commonality than difference. Many of us (Keith, 1999; Lohman, 2005; Willingham, Lewis, Morgan & Ramsit, 1990) have investigated the predictors of academic success in different ethnic groups. We take large data sets and extract the test scores for all of the Black, Hispanic, or Asian-American children. We then look at the ability variables that predict academic success for each group of children. What we have found is that the ability and achievement variables that best predict academic success for minority students are the same as those that best predict academic success in non-minority children. In academic domains, these predictive variables are prior achievement in the domain, the ability to reason in the symbol systems (language, numbers, music notation, etc.) used to communicate new knowledge, interest in the domain, and the ability to persist in striving for excellence in it. This means that if the goal is to identify those minority students most likely to excel in mathematics, then one should find the minority students who currently display the best mathematics achievement, who score the highest on tests of quantitative and nonverbal reasoning, who express an interest in mathematics, and whom teachers rate as being motivated and persistent. For success in verbal domains, the best predictors are current achievement in those domains, verbal reasoning abilities in the language(s) of instruction, interest, and motivation.

However, some schools do not measure these characteristics. Instead, they rely on nonverbal tests – such as the nonverbal battery of the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001a) or, more commonly, tests like the Progressive Matrices (Raven, Court, & Raven, 1983). There are several reasons for this. First, children who are not native speakers of English are clearly at a disadvantage on tests that use English. Nonverbal tests reduce the influence of language and therefore increase the number of bilingual and ELL students who are included in the program. Second, it is sometimes asserted (e.g., Naglieri & Ronning, 2000) that nonverbal, figural reasoning tests predict academic achievement as well as tests with verbal or quantitative content. Others consider nonverbal tests exchangeable with verbal and/or quantitative ability tests because they are good measures of the general ability factor Spearman called *g*. Because of this, they believe that this makes them equally good as selection tests for everyone. Third, most people believe that an unbiased ability test would measure something like the innate potential or capacity of the learner. It is easier to believe that this might be true for a test that uses spatial figures than for a test that uses words, numbers, or other symbols. Fourth, the merits of nonverbal tests have been exaggerated by some.

Construct–Relevant and Construct-Irrelevant Score Variation

An important source of test score invalidity is the presence of unwanted sources of difficulty. These introduce what psychologists call "construct-irrelevant" variation into the scores. On a test of mathematical skills, for example, the ability to understand the directions in English is unrelated to the construct of interest—i.e., the ability to solve math problems. By reducing or eliminating this source of difficulty, one can better compare the scores of all children to the same set of norms or standards. However, the ability to understand the English language can also be an important aspect of the construct one hopes to measure. In this case, the variation would be construct-relevant rather than construct-irrelevant. For example, how well an ELL child understands ideas that are communicated in English can be a critical aptitude for success in a class in which English is the language of instruction. More generally, eliminating verbal content may reduce the extent to which the test adequately represents the construct of interest for all children. An enormous amount of our cognitive competence is either mediated by language or is the direct outcome of language acquisition and use. Indeed, the ability to reason depends critically on what we know and on how well we regulate our thinking. Both are rooted in language. Expelling language (and numbers and other symbols) results in a test that captures only a small part of our ability to reason.[2] Additionally, reasoning with spatial figures does not eliminate content specific to the test items. Reasoning with figures is content-specific just as reasoning with words is content-specific. The important question, then, is "What is the construct that we hope to measure with this ability test?"

We have long known that one of the most important aptitudes for academic learning is the ability construct Spearman (1923) called *g*. Spearman believed that virtually all cognitive tasks required *g* to one degree or another. If the variability in scores on a test is represented by a circle, then the correlation between two tests would be indicated by the overlap in the circles that represent them. When tests correlate highly the two circles would overlap much (see Figure 1). The general ability factor *g* would be represented by the overlap among circles for all of the different tests in a battery. Spearman – and most psychologists after him – was concerned primarily with the overlap among tests. In his analyses, the non-overlapping score variation was discarded. This is useful when developing theories about what unobservable traits might cause tests to correlate. But it can mislead practitioners. Those who use test scores get all of the variation – both the shared part and the non-shared part.

Why does this matter? One of the most pervasive misunderstandings in the field is the belief that all measures of general ability (or g) are more or less interchangeable. If one cannot administer a Binet or Wechsler test, then the Raven will measure the same thing.[3] But this is not true. Even though nonverbal, figural reasoning tests such as the Progressive Matrices (Raven, Court, & Raven, 1983) are good measures of *g*, they are not exchangeable with selection tests that use verbal and quantitative content any more than weight and height are interchangeable measures of general physical growth. Indeed, only about half of the variation in scores on the best tests can be attributed to *g*. The remaining half reflects the influence of other cognitive factors, things that are specific to the test and its format, and errors of measurement. *This means that differences between students in the scores that they obtain on a nonverbal test are as likely to be caused by factors other than g as by g.*

Second, as in the example of height and weight, whether these other factors help or hurt when predicting success depends on what exactly one wants to predict. To weigh more could be helpful in football; to be taller could be more advantageous in basketball. Both common sense and careful study show that success in school depends heavily on children's abilities to understand what other people say and to communicate their own thoughts in words. Verbal reasoning abilities are thus critical for success in school in any culture. Indeed, the bilingual child's ability to reason with words in the English language is an excellent predictor of how well he or she will do in schools in which English is the primary language of instruction. This makes good sense psychologically. Anyone who has struggled to understand another language knows full well that the ability to make good inferences about the meaning of unfamiliar words is a constant – not a sometime – activity.

On the other hand, figural reasoning ability is a more distal (and thus relatively poor) predictor of success in academic learning for all ethnic groups – White, Black, Hispanic, and Asian-American. For example, the CogAT Nonverbal score correlates about r = .6 with reading achievement on the ITBS. However, the CogAT Verbal score correlates about r = .8 with reading achievement (Lohman & Hagan, 2002). Although this many not seem like very large difference, it actually makes the verbal reasoning score a much better predictor of success in reading. For example, when looking at the most able students (e.g., top 1 percent), a correlation between tests of r = .6 means that 19 percent of students will score in the top 1 percent on both tests. When the correlation is higher, r =.8, the proportion rises to 38 percent. This means the predictor test correctly identifies gifted students twice as often when the correlation between tests is .8 than when it is .6. Therefore, even small differences in correlation have large impacts on the accuracy of identification of able students. Only about a quarter to a third of those who those obtain scores above the 97th percentile on the nonverbal test are those who currently obtain similarly high scores on achievement tests in mathematics or science or reading or any other academic domain. Selecting students on the basis of their nonverbal test scores thus eliminates the majority of high-achieving students in all ethnic groups

Finally, once one has accounted for the *g* variation in figural reasoning tests, the specific part sometimes shows a *negative* relationship with success in school. In fact, students whose CogAT Nonverbal reasoning scores are significantly higher than their CogAT Verbal and Quantitative reasoning scores actually do *less* well in school than students who show a relative weakness on figural reasoning tests.[4] In other words, high scores on a nonverbal test can indicate an inaptitude for schooling.

The Importance of Multiple Perspectives

Non-native speakers of English can be disadvantaged on tests that use English. Nonverbal tests reduce (but do not eliminate) the influence of language and therefore increase the number of bilingual and ELL students who are included in the program. Because of this, even those who recognize that expelling verbal and quantitative concepts excludes an enormous amount of cognition often resort to nonverbal tests. The unstated assumption is that all children must be compared to all other children in the nation who are exactly the same age or who are in the same grade in school. This is neither necessary nor desirable.

The appropriateness of the norm or reference group depends on the inference that one wants to make. Inferences about aptitude usually require different norm groups than inferences about level of accomplishment. The surest indicator of aptitude for anything is the observation that the person learns in a few trials what it takes other people many trials to learn. This means that inferences about aptitude are defensible only when one has controlled for opportunity to learn.

On ability tests, opportunity to learn is approximated by the child's age. We estimate the 6 year–3 month old child's aptitude for learning those skills that collectively define the construct of intelligence by how well she performs compared to other children who have been living in the culture for 6 years and 3 months. Changing this reference group by a few months changes the estimate of the child's learning ability. 6 years 3 months is an inappropriate reference group if the child has not lived in the culture for this long. For example, the current level of competence of a bilingual child in using the English language might be only at the class average. But if she has had much less opportunity to learn English than the other children this could indicate a remarkable aptitude for learning English. The only way to know this would be to compare her performance to that of other bilingual children who have had roughly similar learning opportunities. In the case of most skills, one can do quite well by comparing each child to others of approximately the same age (or grade) who have had little, some, or much experience in the domain. Two or three levels of experience will do. The tradeoff here is between making precise statements about a student's rank within the wrong norm group and less precise statements about her rank within the right – or at least a better – norm group.

Why is this not done more routinely? There are several reasons. First, those who come from a tradition in which each child is assessed individually have no easy way of creating these norms for the local population or opportunity-to-learn subgroups within that population. This is not the case for group-administered tests. If all the children in a particular grade in a school district are administered a test, one can easily look not only at the child's rank on national norms, but also at her rank compared to local population, and even to subgroups within the local population. Second, most people believe that ability tests measure – or ought to measure – innate ability. This makes irrelevant any discussion of opportunity to learn. Third, it is administratively convenient to use a single norm group.

A caution or clarification. Knowing that a child is doing well when compared to others who also have had limited opportunities is useful for making inferences about aptitude but much less helpful when making inferences about the child's current educational needs. To know what instruction is appropriate for a student, we also need to know her level of achievement when compared to others in the classes in which she will participate. These inferences typically require common norms or standards. The most sensible policy is to get *multiple perspectives* on the child by comparing the child's test score to several different norm groups: the nation, the local population (e.g., the district or school), and opportunity to learn subgroup (e.g., ELL versus native speakers in the class). Procedures for doing this are outlined below.

Innate Ability

Novices in any field tend to judge things by their appearances. This holds true for judgments about what tests measure. Figural reasoning tests appear to measure something that is less the product of experience than, say, a series completion task that uses numbers or letters of the alphabet. These in turn would be judged as less influenced by experience than a vocabulary test that required a clear understanding of the meanings of commonly used words. But appearances can be misleading. The relative contribution of heredity and environment is the same for all three. Indeed, it is the figural reasoning test that is most subject to practice effects and which has shown the greatest increase in scores over the past 50 years (Flynn, 1999)[5].

There are no culture-free or culture-fair tests (Anastasi & Urbina, 1997; Scarr, 1994). However, this is not the message we want to hear. Good people want to believe that if we could just get it right, we could in fact eliminate bias and then measure innate ability in a way uncluttered by experience or education. But we cannot measure innate ability. All ability tests measure developed abilities; they are really just special kinds of achievement tests.

Instructional Needs

Another myth about testing and giftedness is that once students have been identified as “gifted,” all will be ready for advanced instruction of one sort or another. However, students who obtain exceptionally high scores on selection tests will often have markedly different instructional needs. When students are selected on aptitude as well as accomplishment, some who show stellar achievement in the domain will be ready for advanced instruction covering content that is normally presented to students two or three grades older. Other students who show great promise but not outstanding prior achievement will have the ability to learn more rapidly than their peers. They may be ready for accelerated instruction, perhaps covering a year’s worth of content in one semester. However, these students will generally not succeed if suddenly placed in a class covering content far beyond their current knowledge. Therefore, although both the students with strong accomplishment and those with strong potential need special instruction, they often need different forms of special instruction.

Nonverbal Tests as the Panacea?

As any one who works in education knows, differences between under-represented minority and majority students on both achievement and ability tests are substantial – typically in the range of a half to a full standard deviation. Further, even small differences at the mean translate into substantial differences at the tails of the distribution. Therefore, the claim that one nonverbal test identified equal proportions of high-scoring Black, Hispanic, and White students surprised many (Naglieri & Ford, 2003). However, other investigators have been unable to replicate this finding – either with Black students (Shaunessy, Karnes, & Cobb, 2004; Stephens, Kiger, Karnes, & Whorton, 1999), Hispanic students (Lewis, 2001), or other groups of ELL students.[6] For example, the St. Paul Public Schools have administered the Nagileri Nonverbal Ability Test (NNAT; Naglieri, 1997) to all kindergarten and second grade students for several years now. In an evaluation of the program, Drake (2005) reported that there were approximately 3.5 times as many White as Hispanic students and 6.2 times as many White as Black students who scored above 130 on the NNAT. Lowering the cut score to 120 reduced these disparities somewhat, but resulted in the labeling of 38.7% of the Caucasian kindergarten children and 33.3 % of the Caucasian second graders as "gifted."[7]

As this study and many others show, nonverbal tests such as the NNAT reduce but by no means eliminate differences between ELL and native speakers of the English language. More importantly, they miss most of those students in all ethnic groups who currently show the highest levels of academic achievement. It seems odd that in the rush to be fair schools would implement systems for identifying the most academically talented students that will actually eliminate most of them.

The Process

We have argued that the best way to identify students who are likely to excel in particular domains is to measure the aptitudes that are most needed for successful learning in those domains under particular instructional arrangements. In this section, we show how this can be done. We also show how scores on nonverbal tests can contribute to the identification process.

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

The second procedure is simpler. It shows how to combine information from a nationally (or locally) normed ability test with teacher ratings of only those students who are nominated for the TAG program. It can be used to help make decisions about which students are most likely to profit from some form of acceleration and which might best be served by enrichment or additional instruction at grade level.

The third procedure combines elements of the first two procedures. It shows how best to combine scores on ability and achievement tests. Then these composite scores are compared with teacher ratings to inform decisions about acceleration or enrichment for those who are identified as academically talented.

These procedures were developed using the CogAT – Form 6 (Lohman & Hagan, 2001a), the Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001) and the Scales for Rating the Behavioral Characteristics of Superior Students (SRBCSS; Renzulli, White, Callahan, Hartman & Westberg, 2002). Decisions about educational programming – especially whole grade acceleration – would require additional information, such as some measure of the student's interests, social skills, anxiety, and other characteristics that would be expected to influence the probability of successful learning in the different educational placement options under consideration (Assouline, Colangelo, Lupkowski-Shoplik, Lipscomb, & Forstadt, 2003). These measures are not discussed here. Rather, we focus on a principled way to identify those who are the best candidates for enrichment or acceleration.

The Cognitive Abilities Test

Since the procedures here are based on the CogAT, it is important to understand something about this test. The CogAT is a group-administered series of reasoning tests that are organized into three different batteries that measure verbal reasoning, quantitative reasoning, and nonverbal reasoning. These correspond with the (verbal) sequential reasoning, quantitative reasoning, and (figural) inductive reasoning abilities that Carroll (1993) identified as the three main constituents of the general fluid ability (Gf) factor.[8] On the Primary Edition of the test (grades K – 2), each battery has two untimed tests. On the Multilevel Edition (grades 3-12), each battery has three tests. Using two or three different tests to measure each of the three reasoning abilities reduces the confounding effects of test format that occur when all items follow the same format, thereby increasing both the construct validity and reliability of the three reasoning scores. In addition to the scores for each of the three batteries, an overall composite score is routinely reported as well. Partial composite scores (e.g., a Quantitative-Nonverbal or QN Composite) are also available from the publisher (or can be computed by test users). As will be seen, the QN partial composite is an important component of one of the identification procedures that we describe. The overall composite score is not used and generally should not be used in the identification of gifted children (Thorndike & Hagen, 1978; 1987; 1993; Lohman & Hagan, 2001b,c).

In addition to its broad coverage, there are several unique features of CogAT that can assist users in avoiding misuses of test scores. For example, the pattern of each student's responses is analyzed to detect inconsistencies. A warning is printed if a student misses many easy items but answers difficult items correctly, or if she performs much better on one subtest than on the other two subtests in a battery. This can happen, for example, if the student skips one row on the answer sheet or misunderstands the directions for one of the subtests.[9] Extensive interpretive manuals – both print and on-line – provide explicit guidance on how to interpret each student's score profile. Importantly, only about 30 – 40 percent of the students at all score levels obtain scores on the three batteries that do not differ significantly. This means that the majority of students are not well described by their overall composite or g scores. Instead, they show relative strengths or weaknesses in their reasoning abilities. In fact, profiles that show a markedly lower score on one of the three batteries are much more common among the most able students than among average ability students (see Lohman & Hagen, 2001c).

Procedure 1. Multiple Norm Groups

Detailed directions for using this method are provided in Lohman (2006a) and in the sample data set that accompanies that monograph. Here are the steps for using only one test score. Examples in the sample data set that is available on the website show how to combine scores on two variables (such as mathematics achievement and quantitative reasoning abilities). This is important because identification procedures that average ability and achievement scores for particular domains better identify not only those who currently excel, but also those who are most likely to continue to excel (see Lohman & Korb, 2006).

Step 1 – Preparing the data. Get the required data into a spreadsheet. For each student, this would include the student's name or ID, an opportunity to learn index (such as ELL status), national percentile ranks (PRs) or other norm-referenced test scores. On CogAT, these would be Standard Age Scores (SAS) for one or more of the test batteries.

Step 2 – Getting local ranks. Sort the data by percentile ranks (or SAS scores). This will provide local ranks.[10] Those with highest scores will be at the top of the list. Local score distributions generally provide a better way to determine which students are most likely to be mismatched with the instruction they are receiving than will national norms. They also make it much easier to identify a relatively consistent number of students across years.

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

Procedure 2. Using Ability Test Scores and Teacher Ratings

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

Although the procedure can be used when test scores are available for all students in the local population, it is particularly helpful when only some students are rated by their teachers or are administered an ability test. Whenever only a portion of the population is tested, it is important to test a much broader segment of that population than just those students who score highly on an achievement test or who are nominated by their teachers. Testing only those who meet the desired criterion on an achievement test (e.g., 97th PR) will eliminate the majority students who score higher on the ability test than on the achievement test. Testing more children is much easier to do when using a group-administered test rather than an individually-administered test. A reasonable rule is to test every child who scores above average on one or more of the major batteries of the achievement test (typically Reading Total and Math Total for elementary students). If this is done, it is often more administratively convenient to administer the ability test to all children. Classroom teachers are most likely to go along with this procedure if the ability test does more than identify the most able students but also provides information that they can use to help all students learn.[11]

How best to combine scores from the three *CogAT* batteries when predicting academic success is well documented in the research literature. Importantly, the weights that should be applied to each test battery in making these predictions are the same for all ethnic groups that have been studied (Lohman, 2005). Competence in a broad range of verbal domains (e.g., reading comprehension and literary skills) is best predicted by the CogAT Verbal SAS score. On the other hand, success in mathematics and domains of study that demand quantitative thinking is best predicted by a combination of the CogAT Quantitative and Nonverbal Reasoning Batteries. Further, differences between ethnic groups are often about the same on these two batteries, and so any advantage that might accrue from reduced mean score differences on the Nonverbal Battery are not compromised by combining those scores with scores on the Quantitative Battery. However, adding quantitative reasoning to the mix substantially improves the ability of the test to identify the best students. Therefore, we recommend using the Verbal Battery SAS score and the Quantitative-Nonverbal (QN) *Composite* SAS score to guide admissions decisions.[12] Students should be considered for admission if they obtain either a high Verbal SAS or a high QN Composite SAS.

Research with the *SRBCSS* shows that each of its three main scales provides unique information. Therefore, the decision rule we suggest is that teacher ratings be considered high if any one of the three ratings is high.[13]

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

The identification scheme is shown in Figure 2. The vertical dimension distinguishes children who exhibit superior reasoning abilities from those who exhibit above average reasoning abilities. We have set the cut scores as scoring at or above the 97th national percentile rank or at or above the 80th national percentile rank on either verbal reasoning or quantitative-nonverbal reasoning. These criteria are commonly used in gifted programs. However, other cut scores could be used in order to identify a particular percentage of the applicant pool. This is a relatively easy way to bring local norms into the picture. For example, one could set the criteria as the 97th and 80th local percentile ranks to identify students whose abilities are well above those of their classmates. In some schools, these students would have much higher national percentile ranks whereas in other schools they would have much lower national percentile ranks. Similarly, the horizontal dimension distinguishes between children who, when compared to other children nominated for the program, obtain above average teacher ratings and students who obtain average or below average teacher ratings. Note that, for ratings, the average is computed only on the subset of the student population who are nominated for inclusion in the program. Combining these two criteria gives four categories of assessment results.

Children in Category I exhibit superior reasoning abilities on *CogAT* and are rated as highly capable, motivated, or creative by their teachers. Children in Category II also exhibit superior reasoning abilities but, when compared to other children who were nominated, are not rated as highly by their teachers on any one of the three major scales of the SRBCSS. Programs that follow a traditional identification scheme (e.g., self-contained classrooms or schools) would accept children in Categories I. Some would aso accept children in Category II, given the difficulty of defending rejections on the basis of low teacher ratings. However, the progress of children in Category II should be monitored more closely. Children in Category III exhibit somewhat lower but strong reasoning abilities (80th to 96th PR) on *CogAT*, and are rated as highly capable, motivated, or creative by their teachers. These children would be included in school-wide enrichment programs that aim to serve a broader range of children. Schools that serve mainly poor and low-achieving students would find that many of their best students would fall in this category, especially when using national rather than local test norms. Combining test scores and ratings in this way would enable these schools to identify the students most likely to benefit from curriculum compacting or enrichment programs, including instruction at a higher level than that received by most other students in the school. Finally, children in category IV exhibit good but not exceptional reasoning abilities (between 80th and 96th PR), and are not rated as unusually capable, motivated, or creative by their teachers. Although good students, these children would not be provided with special programming on the basis of either their *CogAT* scores or teacher ratings. However, if rank within opportunity-to-learn group is high, then some form of special assistance could be provided.

Procedure 3. Using ability scores, achievement scores, and teacher ratings. Although either of the procedures we have described can be used to improve the identification process, our preferred procedure would use achievement in addition to ability test scores. This is because the best prediction of subsequent success in school is given not by CogAT scores alone, or achievement test scores alone, but by the combination of CogAT scores and current achievement in the domain. Further, the best way to combine these scores is to average them.[14] An easy way to average scores is to get the data for both ability and achievement test scores into the spreadsheet, put them on the same score scale, and then average them.[15] Then either procedure 1 or 2 can be followed using these averaged scores. Directions for doing this are given in the sample data set on the web.

Uses of Nonverbal Tests in Screening for Gifted Students

How can nonverbal ability tests assist in identifying academically talented students? They are most useful as supplementary measures for predicting success in mathematics and technical domains (e.g., computer programming). Nonverbal tests should never be used to screen all children. This would be like measuring height when what we really need is weight. Height and weight are positively correlated. We can predict weight from height, but only with much error. It is not fairer to measure height for everyone just because we find it difficult to measure the weight for some. Rather, we should use predicted weight only when we cannot actually weigh people. *The critical point, however, is not to confuse a higher average nonverbal score with better assessment of the relevant aptitudes.* Put differently, the nonverbal, figural reasoning test may appear to reduce bias, but when used alone, it actually increases bias by failing to select those most likely to profit from advanced instruction.

High scores on figural reasoning tests tell us that students can reason well about problems that make only the most elementary demands on their verbal and quantitative development. This is extremely useful when one must determine whether a child suffers from a general cognitive impairment. From the early form boards of Itard to the contemporary nonverbal ability tests like the Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998), nonverbal tests have assisted psychologists in making inferences about the abilities of children who have hearing or speech problems or who do not speak the language of the examiner. But it is one thing to make inferences about basic mental competence and another to infer readiness to profit from advanced instruction. Absent information on verbal skills in the language of instruction, scores on a nonverbal test tell little about whether children will succeed in classes conducted in Spanish, Japanese, or any other language. More importantly, even *within* the population of native speakers, those students with the highest nonverbal reasoning scores are usually *not* the students who are most likely to show high levels of achievement in the classroom. In fact, nonverbal tests sometimes have a negative influence on the prediction of success in domains that require verbal fluency.

Those students with the highest academic achievement in specific domains and those who reason best in the symbol systems used to communicate new knowledge in those domains are the ones most likely to achieve at a higher level. Therefore, high nonverbal scores should qualify students for acceleration or enrichment *only if* the scores are accompanied by (a) evidence of reasonably high accomplishment in the academic domain in which accelerated instruction or enrichment is offered and (b) evidence that the student's verbal or quantitative reasoning abilities are also high *relative to other children who have had similar opportunities to develop these abilities.* Most schools have this evidence for achievement, and those that administer ability tests such as CogAT that appraise verbal and quantitative reasoning in addition to nonverbal reasoning have the corresponding evidence for ability as well. For these schools, procedures like those outlined here, combining evidence of current achievement, reasoning abilities, and teacher ratings can help increase the diversity of gifted programs while also identifying the students in all ethnic groups most likely to benefit from special instruction.

References

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

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

Bracken, B. A., & McCallum, R. A. (1998). *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside.

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

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

Drake, S. (2005) *Gifted services identification report*. Office of Research & Development, Department of Research, Evaluation, and Assessment, St. Paul Public Schools St. Paul, MN.

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

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

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

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

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

Lohman, D. F. (2006a*). Identifying academically talented minority students*. (RM). Storrs, CT: National Research Center on the Gifted and Talented, University of Connecticut. [Sample data set and a draft of this monograph are available at http://faculty.education.uiowa.edu/dlohman/]

Lohman, D. F. (2006b). *Identifying academically gifted children in a linguistically and culturally diverse society*. Keynote Presentation at the Eighth Biennial Henry B. & Jocelyn Wallace National Research Symposium on Talent Development, Iowa City.

Lohman, D. F. (2006c). *Fluid abilities are more than figural reasoning ability*. Paper presented at APA, New Orleans,

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. (2001c). *Cognitive Abilities Test (Form 6): Interpretive guide for school administrators.* 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.

Lorge, I., & Thorndike, R. L. (1954). *The Lorge-Thorndike Intelligence Tests.* Boston: Houghton-Mifflin Co.

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 Stanford Achievement Test (SAT) reading achievement. *Journal of Psychoeducational Assessment, 18,* 230-239.

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

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

Renzulli, J. S., Smith, L. H., White, A. J., Callahan, C. M., Hartman, R. K., & Westberg, K. L. (2002). *Scales for Rating the Behavioral Characteristics of Superior Students*. Mansfield Center, CT: Creative Learning Press.

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.

Spearman, C. E. (1923). *The nature of intelligence and the principles of cognition*. London: Macmillan.

Stephens, K., Kiger, L., Karnes, F. A., & Whorton, J. E. (1999). Use of nonverbal measures of intelligence in identification of culturally diverse gifted students in rural areas. *Perceptual and Motor Skills, 88,* 793-796.

Thorndike, R. L., & Hagen, E. (1978). *Cognitive Abilities Test (Form 3)*. New York: Houghton Mifflin.

Thorndike, R. L., & Hagen, E. (1987). *Cognitive Abilities Test (Form 4)*. Chicago: Riverside.

Thorndike, R. L., & Hagen, E. (1993). *Cognitive Abilities Test (Form 5)*. Chicago: Riverside.

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

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children.* New York: The Psychological Corporation.

Figure Captions.

Figure 1. Schematic diagram of the correlations among three tests. Correlation between any two tests is indicated by the extent to which their circles overlap. The shared scores variance for all three tests is given by the small area at the center of the diagram.

Figure 2. A method for combining ability test scores on two dimensions (Verbal SAS or Quantitative-Nonverbal Composite SAS) and teacher ratings on three scales (Learning Ability, Motivation, and Creativity). National or local percentile ranks are used for ability test scores. However, teacher ratings are obtained only on those students who are nominated for the program.

Figure 1

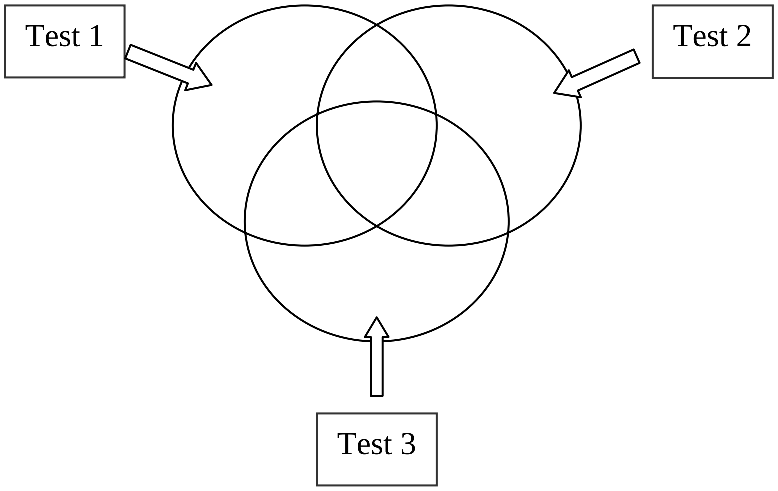

Figure 2

Teacher Rating on Learning Ability, Motivation, or Creativity

		Below average teacher ratings	Above average teacher ratings
***CogAT* Verbal or Quantitative-Nonverbal Reasoning**	>97th PR	II	I
	>80th PR	IV	III

Footnotes

[1] For a summary of Snow’s contributions and a more in-depth discussion of the concept of aptitude, see Corno, Cronbach, Kuppermitz, Lohman, Porteus, & Talbert, 2002).

[2] It is sometimes claimed that one can measure verbal concepts with pictures as well as one can with words. This is not true. Consider even the simplest verbal analogy or verbal classifications items on CogAT, for example. Most cannot be presented in pictures and, if they can be presented, the item would be ambiguous or would represent a much less sophisticated concept. Consider kinship relations. One of the earliest classification schemes that all children learn is for relationships within the family. Consider the analogy *mother is to father as grandmother is to grandfather*. How could one draw a pictorial analogy that measured these kinship relationships rather than the more generic concept *man is to woman as old man is to old woman*?

[3] There are many different kinds of nonverbal tests. In this chapter, we focus on group-administered, figural reasoning tests such as the Progressive Matrices (Raven, Court, & Raven, 1983). Although individually administered nonverbal tests such as the UNIT (Bracken & McCallum, 1998) include a broader range of test-tasks, they typically predict success in school about as well as the group-administered, figural reasoning tests.

[4] The pattern of higher verbal and quantitative scores than nonverbal scores is particularly common among Black students (Lohman, 2005). This means that screening students on the basis of their nonverbal test scores will not only eliminate many of the most academically capable Black students, it will also include even more students who seem to find conventional schooling particularly uncongenial to their preferred ways of thinking.

[5] Most people are unaware of the extent to which norms on ability and achievement test have changed over the past generation. IQ scores have increased at the rate of approximately 3 points per decade. Changes have been even larger on nonverbal reasoning tests making these norms prone to overestimating the ability of test-takers, especially when such tests are used only for minority groups (Flynn, 1999).

[6] Basically, the data were altered to fit the conclusions. For an explanation of how this was done, see Lohman (2006b).

[7] If the two score distributions have similar means, then lowering the cut score will always increase the proportion of students from the lower scoring group who are selected. As this example indicates, however, there is a cost to be paid in the number of students in both groups who exceed the cut score.

[8] Since the fluid reasoning factor (Gf) is defined by the common variation in all three of these abilities, measuring only one reasoning primary under-represents the construct. Elsewhere I show that this has importantly distorted research on the investment theory of aptitude (see Lohman, 2006c).

[9] In fact, these procedures were developed after reviewing the case of a gifted boy in Seattle who made this mistake in using the test answer sheet. See the case study of Maxwell in Lohman & Hagan (2001b).

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

[11] As far was we know, CogAT (Form 6) is the only ability test that provides this sort of information. Score reports contain a profile index that summarizes the level and pattern of scores across the three batteries for every student. This profile is linked to specific suggestions on how to adapt instruction to improve the likelihood that the student will learn. A free interpretive guide that provides more general advice on adapting instruction to individual differences is also available on the web at www.cogat.com.

[12] This composite score can be obtained from the publisher when requesting score reports or computed by averaging the USS scores for the Quantitative Battery and the Nonverbal Battery and then looking up the corresponding SAS and PR scores. Note that simply averaging SAS or PR scores on the separate batteries will not give the proper composite score.

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

[14] Requiring a high score on both tests (or on either test) gives less satisfactory results. For an explanation, see Lohman and Korb (2006).

[15] Averaging test scores that are not on the same scale weighs the test with the greater variability. Percentile ranks (PR) should generally not be averaged. The PR of the average scale score is generally not the same as the average PR.