Identifying academically gifted children in a linguistically and culturally diverse society

David F. Lohman

University of Iowa

May 22, 2006

Invited presentation at the Eight Biennial Henry B. & Jocelyn Wallace National Research Symposium on Talent Development, University of Iowa, Iowa City

I must thank Nick and Susan for inviting me to speak at this symposium. I hope that you have been enjoying your time here. But I also hope that you did not feel too welcome at the luncheon buffet or else my task will be harder that it might otherwise have been! These sorts of presentations are never easy for me. I fret and stew over what to say and how to say it. I worry that – when I finally stand here six feet above contradiction – I will stumble over my words. When I was a boy, I often stuttered terribly – especially when I had to speak in front of my classmates. It was the custom then for children to stand when called upon to answer a question. Sometimes this was a less-than-pleasant experience. I was not called upon that often, though, because there were 56 other children in my first grade class. I did not know that this was unusual. Rather, it was the fact that there were two other boys named David in the class that struck me as a coincidence of cosmic importance.

It was in this class, though, that I came to understand something about my own precocity and the dangers that attended it. The 57 squirming first-graders were managed by one diminutive but remarkable nun – Sister Zacheus. But even she could not teach that many children to read at one time. So she divided us into four groups, three of which were lead by a student whose reading skills were at least one month ahead of the other children in the class and the fourth group by Sr. Zacheus. I was picked to lead one group. It was my first great academic honor. I carefully observed Sr. Zacheus run her group and did the same things in my group. We went around the circle, each child reading a sentence while I monitored their performance, carefully pointing to the words in my basal reader so that I did not lose my place.

On the first day, we got about half-way around the circle when one little boy got stuck on a word. Taking my cue from Sister Zacheus, I stammered "Tommy, don't you know what this says?" Tommy shook his head… probably as much befuddled by me as by the unfamiliar word. And so I turned in my seat to print the word on the blackboard behind me – just Sr. Zacheus did in her group. It was then that I realized to my horror that I did not know how to write. Years later when I read the story of Icarus it had a strange, intuitive appeal. Thus began my career as a small group leader. My graduate students will probably affirm that my performance hasn't gotten a whole lot better over the years.

* *

My theme today is that the procedures for identifying gifted students used in many schools exclude many of the most academically talented minority students. In this talk I will outline a different approach for identifying these students. Four principles guide the approach. First, to identify the right students one must measure the right aptitudes. This requires that one first specify the kind of expertise that is to be developed and the requirements of the educational systems that are available to develop it. Second, different inferences from test scores require different comparison or norm groups. Common norms and standards are appropriate for inferences about academic competence. However, inferences about aptitude require comparisons to others who have had similar opportunities to acquire the abilities measured by a test. Using national norms to estimate the academic talents of all students leads either to the use of tests that are inferior measures of academic aptitude or to the identification of very few minority students. Third, 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. Fourth, rethinking the goals of TAG programs and the range of students and services that they provide could make these programs central to the school's broader mission.

Measure the Right Aptitudes

I had quite a shock last month. I was going through a large pile of unopened mail on my desk when I came across a glossy, 4-page newsletter from AERA. The topic of the newsletter was foreign language instruction. Near the bottom of page 2 my eye caught the following heading: “Can everyone learn foreign languages well?” Then the sentence “A student’s aptitude … can be a key factor in his or her foreign language learning.” There followed two paragraphs that described how foreign language aptitude could be measured and why it was so important. I could not believe it. There it was, in broad daylight, an acknowledgement that not all students can learn all things with equal facility, and that we have devised good ways of measuring the cognitive aptitudes that are required for learning – or at least for learning foreign languages.

I was surprised because, for many years I watched as my mentor – the late Richard Snow – tried to convince educational researchers of the importance of the concept of aptitude. Many seemed to like his message, but few had the temerity to use the *A* word themselves. Snow believed that the concept of aptitude was central in all of psychology. Aptitude, he said, was not only education’s most important raw material; it was also its most important product. Indeed, he viewed education as a systematic aptitude development program. A good education enhances the students’ readiness for new challenges. But he defined the concept of aptitude much more broadly than others. As he used the term, aptitude implies a propensity or readiness or aptness for learning or performing in a particular situation. The attainment of expertise in any domain requires many different kinds of personal resources – some cognitive, some affective, and some conative. And the particular mix of aptitudes required for success varies systematically across the school years. Indeed, one of the most important features of an aptitude perspective is that it helps one go beyond simplistic talent identification systems that ignore interest, motivation, perseverance, anxiety, or even accumulated knowledge and skill in a domain. Aptitude cannot be understood apart from either the kind of learning that must occur or the context 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. Next, one must understand the demands and affordances of educational systems that students must negotiate if they are to attain the desired end state. Thus, an aptitude perspective offers a principled way to study how different kinds of educational systems elicit or require different personal resources. Of course, most talent identification systems are far more restricted. Or, if they collect such information, they have no empirically substantiated way of combining it to identify those most likely to attain expertise.

Fine, you say, but what does this have to do with the identification of academically talented minority students? A lot, actually. I and others (e.g. Tim Keith at U Texas) 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 children or the Hispanic children or Asian-American children. We then look at the ability variables that best predict academic success of these children at different grades. What we have found is that the best predictors of academic success for minority students are the same as those that best predict academic success in non-minority children.

This means that, if you want to identify those minority students most likely to excel in mathematics, you should look first at the students' current mathematics achievement, secondarily at their ability to reason quantitatively, and third at other aptitude variables that add to the prediction of success in mathematics. For success in verbal domains, the best predictors are current achievement in those domains and verbal reasoning abilities in the language(s) of instruction. However, many schools do not do measure these abilities for English Language Learners or, if they do, they do not use it. Instead, they rely on nonverbal tests – such as the nonverbal battery of CogAT or, more commonly, tests like the Raven.

There are several reasons for this:

1. Nonverbal tests are often good measures of *g*. Because of this, some professionals and many tests users believe that this makes them equally good as selection tests for everyone.
2. 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 when common norms are used.
3. Wittingly or unwittingly, some continue to mislead users about the merits of these tests.

Let's look more closely at these points.

Are Good Measures of g Exchangeable

We have long known that one of the most important aptitudes for academic learning is some measure of the factor Spearman called *g*. Spearman believed that virtually all cognitive tasks required *g* to one degree or another. If the variability in scores on a task is represented by an oval, then *g* would be represented by the overlap among the ovals for different tasks. Spearman – and most psychologists after him – was concerned the overlap. In their factor analyses, the non-overlapping score variation is discarded. This is useful for theory but can mislead practitioners. Although many different tasks are good measures of *g*, they are not exchangeable as selection tests. Those who use test scores get all of the score variation – both the *g* part and the non-*g* part.

Why does this matter? One of the most pervasive misunderstandings in the field is the belief that all measures of *g* are more or less exchangeable. If one cannot administer a Binet test, then the UNIT will measure the same thing. But this is not true. Why this is so may be easier to see in the domain of physical skills.

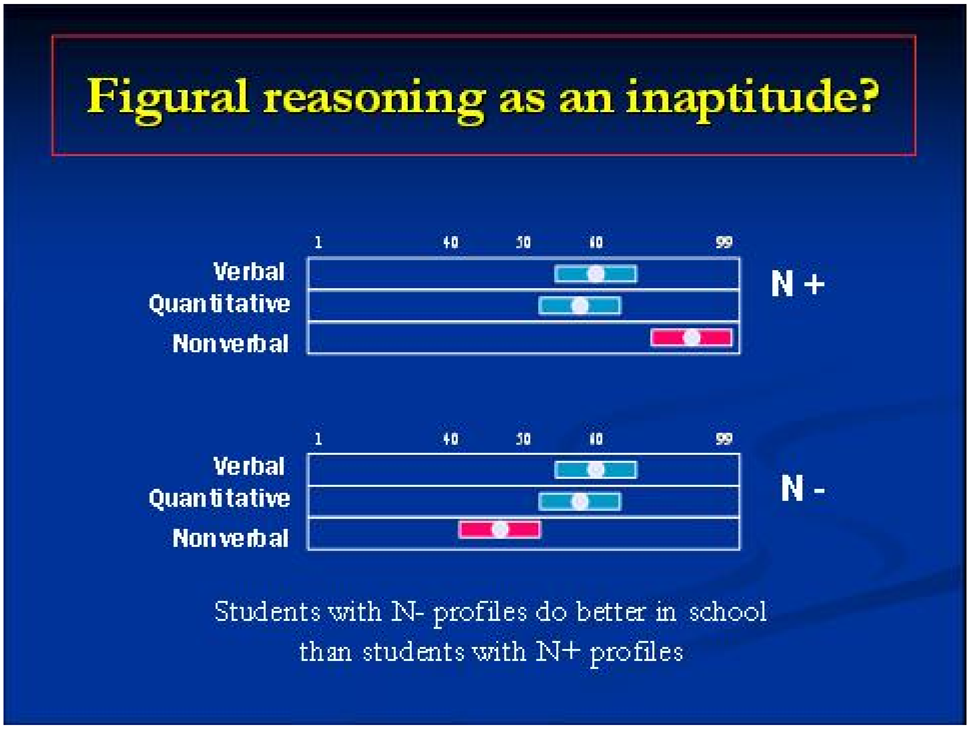
Suppose that, instead of identifying academically gifted children, your job was to identify kids most likely to excel in a special basketball program. As a well-trained coach, you know that one of the best predictors of basketball skill is height. And so you routinely screen kids for height. This would be analogous to selecting on the basis of a Full Scale IQ score on the Wechsler or Binet test. You've done this for years, and so you are quite comfortable with the procedure. However, one year some of the girls in the school and their parents complain that you have very few young women on your team. Recently they have threatened to lobby for the dissolution of your basketball program altogether unless you make a better effort to increase the number of girls in your basketball program.

And so you ask an expert in measurement what you might do. "Oh, that's easy" he says, "Height is one component of a more general growth factor" (which he calls gg for short). "The other major component of this gg factor is weight." Indeed both height and weight show equal correlations with the general growth factor. This means that the two measures are statistically interchangeable as measures of gg. Most importantly, weight does indeed show smaller differences between boys and girls.

This new procedure seems odd but when you question it the measurement expert tells you there are many examples that prove his point. (picture of Shaq here) Wow, you say, I guess I have been using the wrong measure! And so you start using a scales rather than a ruler to screen kids. The next year, your basketball team surely looks different than it did before, but not just because there are more girls on the team. Many of the kids – both boys and girls – seem quite unprepared to play basketball. Indeed, some seem to have a distinct inaptitude for basketball or any other sport for that matter. When you ask about this you are told that, all these years, you had mistaken height for real basketball aptitude. This, or something like it, is happening all across the country as programs attempt to increase the diversity of the children that they serve by using nonverbal tests as the primary selection tool for identifying academically gifted ELL and minority students.

It its indeed true that nonverbal, figural reasoning tests like the Raven Matrices or the Nonverbal Battery of CogAT are good measures of *g*. But this does not make them exchangeable with selection tests that use verbal and quantitative content any more that weight and height are exchangeable measures of physical growth. Only about half of the variation in scores on the best nonverbal tests can be attributed to *g*. The other 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 height and weight example, whether these other factors help or hurt depends on the criterion tasks. 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. Any of you who have struggled to understand another language know full well that the ability to make good inferences about the meaning of unfamiliar words is a constant – not a sometime -- activity.



Figural reasoning as an inaptitude?

Students with N- profiles do better in school than students with N+ profiles

On the other hand, figural reasoning ability is at best a distal (and thus comparatively poor) predictor of success in academic learning. Indeed, once one has accounted for the *g* variation in figural reasoning tests, the specific part often shows a negative relationship with success in school. In fact, students whose nonverbal reasoning scores are significantly higher than their verbal and quantitative reasoning scores actually do less well in school than students who show a relative weakness on figural reasoning tests. This holds for all children: White, Black, Hispanic, and Asian-American. Within each of these populations, only about a quarter to a third of those who those obtain the highest scores on a nonverbal test are those who currently display the highest achievement in mathematics, science, reading or any other academic domain.

There are indeed some Shaq O'Neils out there. But most of the heaviest kids are neither the children who currently display the best basketball skills, nor are they the ones most likely to develop such skills. The same applies to figural reasoning tests and success in school. You can identify some of the most academically talented kids using these tests. But you will miss more than you get.

Norms, norms, and more norms

The second reason for using nonverbal tests was the observation that non-native speakers of English are clearly at a disadvantage 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. The unstated assumption here is that all children should be compared to all other children in the nation who are exactly the same age or who in the same grade in school. This is not necessary. Indeed, we constantly shift norm groups when interpreting scores. For CogAT, norm tables shift monthly; for the ITBS, they shift weekly.

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, of course, that inferences about aptitude are defensible only when one has controlled for opportunity to learn.

On intelligence 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 culture for 6 years and 3 months. Changing this reference group by a few months changes the estimate of the child's learning ability. Six years 3 months is clearly an inappropriate reference group if the child has not lived in the culture for 6 years and 3 months. For example, the current level of competence of a bi-lingual child in using the English language might place her at the mean of others in her grade. 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 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 the students rank within the wrong norm group and less precise statements about her rank within the right – or at least better – norm group.

A caution or clarification….

Knowing that I am doing well when compared to others who also have had limited opportunities is useful for making inferences about aptitude but unhelpful when making inferences about my current educational needs. These 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: national, local, and opportunity- to-learn subgroups. I show how to do this. In the monograph recently issued from the National Research Center on the Gifted and Talented.

Why is this not done more routinely? There are several reasons. First, most people are unaware of the extent to which norms on ability and achievement test have changed over the past generation. Second, those who come from a tradition in which each child is assessed individually have no easy way of creating these norms for their 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 school or 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. Third, many erroneously believe that good ability tests measure innate ability, which makes consideration of opportunity to learn irrelevant. Fourth, and probably most importantly, it is administratively convenient to use a single cut score.

Misleading claims

The final reason schools have stated using nonverbal tests to screen kids for gifted programs is harder to talk about. I keep wishing that it would just go away, but it does not. Some of you may know that, several years ago, Jack Naglieri presented a paper at NAGC (and subsequently in many other places) that purported to show that his test – the NNAT – identified equal proportions of high scoring White, Black, and Hispanic students in a large, national sample of school children. He and Donna Ford subsequently published an article in on this in the Gifted Child Quarterly. As any one who works in education knows, differences between under-represented minority and majority students on both achievement and ability tests are enormous – typically in the range of a half to a full standard deviation. Further, as Camilla Benbow pointed out many years ago, even small group differences at the mean translate into substantial differences at the tails of the distribution. Therefore, the claim that any achievement or school ability test gives equal representation of high-scoring Black, Hispanic, and White students is, quite literally, unbelievable.

I did not want to be the one who challenged that claim, though. I knew that some would think it simply sour grapes—I work on a test that does not show these effects. In fact, some have even said this to me. I was also warned challenging these claims would brandish me as an opponent of equal opportunities for minority students. That too has happened. But I also realized that very few people who work in the field of gifted education seemed to have the technical expertise in large scale testing to understand what was going on here. And so I challenged that claim, but was restrained in my comments. I pointed out that

- the numbers did not add up;
- the results were inconsistent not only with every other large data set but also with previously published analyses of the same data set;
- and therefore that the conclusions we not to be trusted.

But I did not explicitly say what I knew – which was that the data had been retroactively fit to the conclusions. I thought that any but the most naïve reader would get the point that the data set had be altered in a serious way I worded the conclusions in this way because I did not want to be confrontational, and I wanted to give the authors a way out of the mess they had created. In my naiveté, I thought they might say something that would allow them to save face and reputation while setting the record straight. I also communicated my concerns privately to the editor, and warned that, if past behavior predicted future behavior, Naglieri would not address the issues that I raised, but would instead attack me.

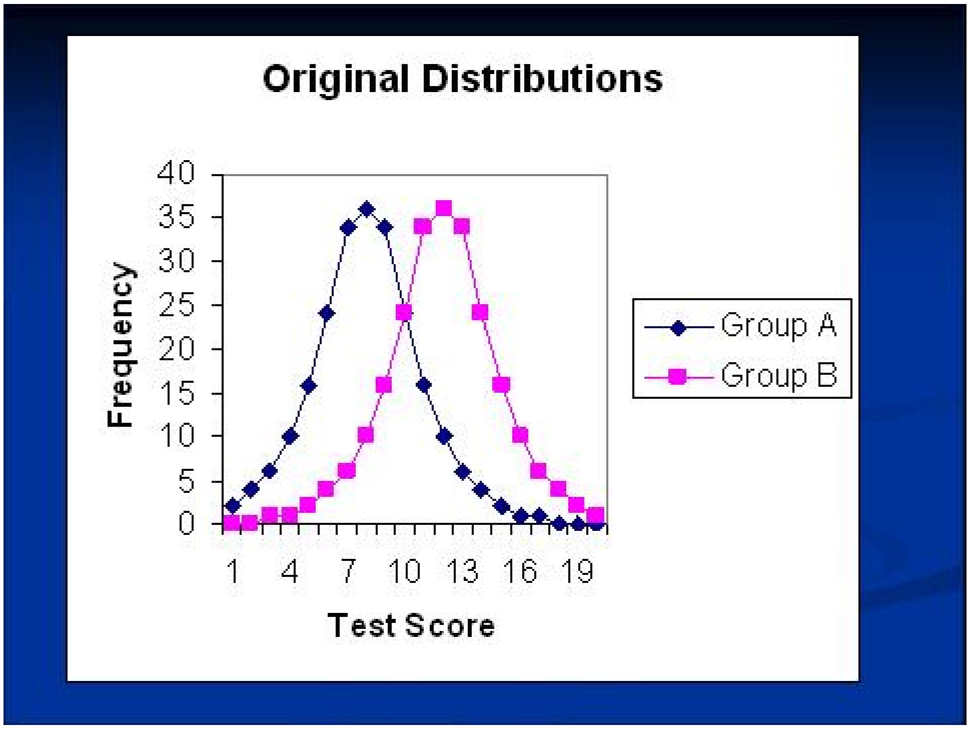
And, indeed, this is what happened. My motives for writing the article were questioned. The CogAT was attacked – most spectacularly with a set of readability numbers that are nothing but random noise. And the authors assumed the tone offended advocates for the downtrodden, while caricaturing me and my work as defending the evil status quo. The only legitimate point that they raised – and illustrated in several pages of text and figures attacking the CogAT and ITBS – was their contention that ability and achievement are independent constructs. The measurement of one, they said, should not be contaminated by the other. I have a very different view – which I have articulated in a paper that will appear in the Fall 2006 edition of the *Roeper Review* and that is on my website.

Needless to say, I was astonished both by what they we allowed to say and, more importantly, what they did not say. There was no admission of re-weighting the data or even of misleading unsuspecting users. Nor was there any explanation for the inconsistencies between their results and previous analyses of the same data set. Burt (because he championed a politically unpopular position) was posthumously pilloried for a decimal point. This is about moving entire distributions amounts that would be classified as very large effects in the experimental literature. The difference between Black & White students on nonverbal tests is about as large as the difference between these groups on measures of academic achievement. For Hispanic students, the differences are reduced but still substantial. If Naglieri had honestly reported ethnic differences on his test, this is what he would be telling potential users.

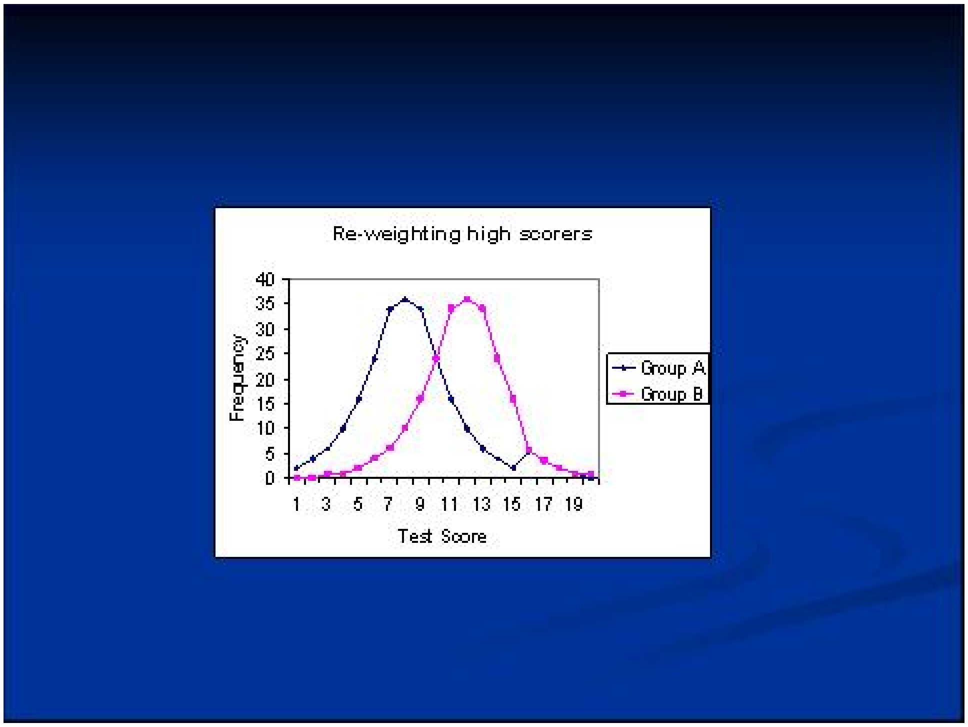
But this is not the message we want to hear. Good people want to believe that if we could just get it right, we could in fact eliminate bias and then measure innate ability in a way un-cluttered by experience or education or anything else. But we cannot measure innate ability. All ability tests measure developed abilities; they are really just special kinds of achievement tests.

I have traversed quite a range of emotions about this. Perhaps it was my fault. Perhaps I should have been more direct. Indeed, most seem to understand the issue as a scholarly dispute of the sort that fills the pages of academic journals. To understand why there is more to it, you must also understand at least in general terms what was done. And so I will outline for you how the data were altered. But before I do so I want to be clear that I do not believe that Donna Ford had anything to do with this. I met her last year at NAGC and found her to be gracious and professional. I believe that she, like many in the field of gifted education, was taken in by this. However, I do not know what her reaction has been to my explanation of how the data were altered. Had someone whom I trusted done this to me and to my reputation, I would not be pleased.

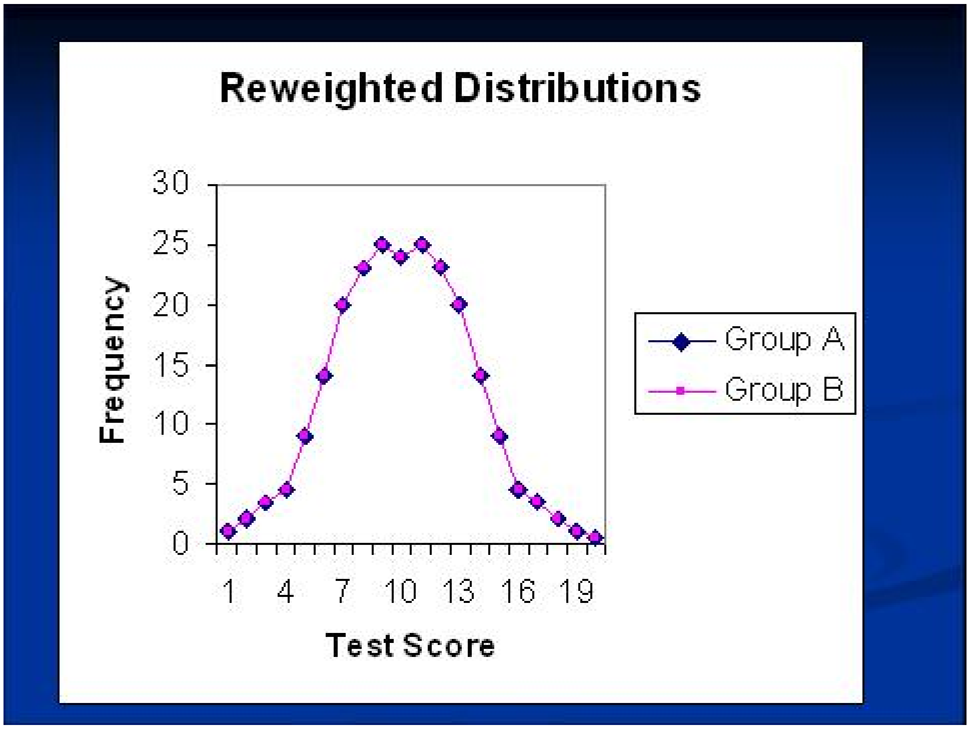
Here is what was done. First, test scores were re-weighted to make the score distributions equal. This guaranteed that there would be equal proportions of students from different ethnic groups. Here is a visual demonstration of the process. The blue distribution is for the lower-scoring minority group, the red for the higher scoring non-minority students.



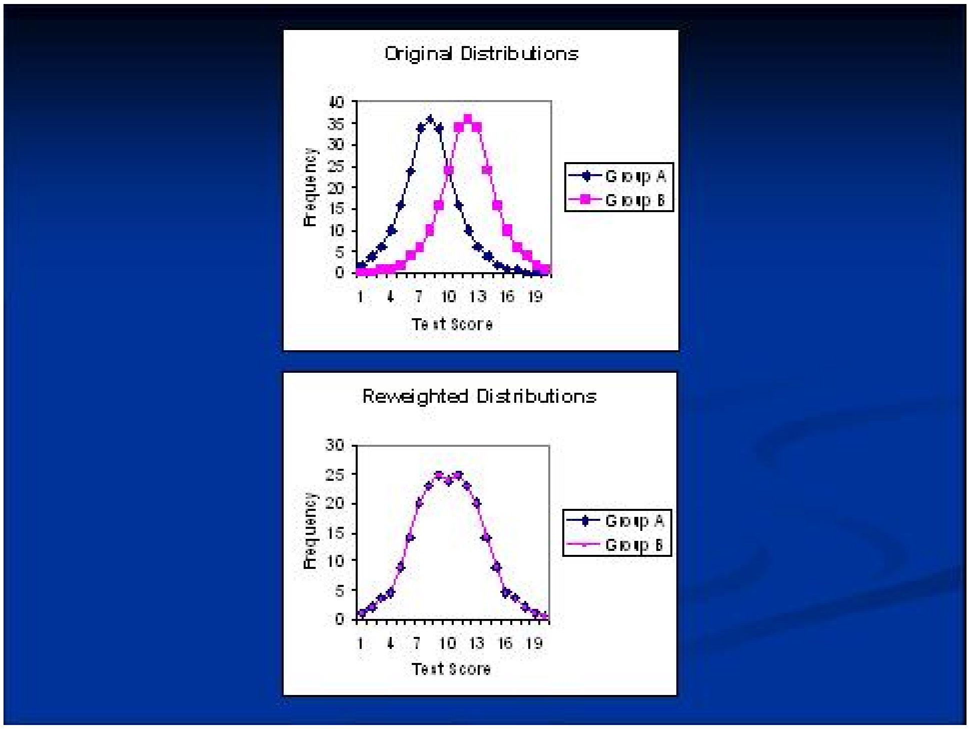
We can make two distributions the same by duplicating or up-weighting the records for high scoring minority students and simultaneously down-weighting or discarding records of high-scoring non-minority students. For example, here I re-weighted the top few categories.



For the last altered value (16), the cases in the lower-scoring group had to be multiplied by approximately 5 and those in the higher-scoring group halved. If you do this systematically, you would get two score distributions that look like this.



Both distributions have the same mean. But the variability of scores is now greater. This was one of the first things that I noticed about the Naglieri data.



He then tallied the frequency of demographic variables for these new, re-weighted data sets. For example, the social class and other demographic characteristics of the students whose scores were up-weighted would now be much more important, and conversely for those students whose scores were down-weighted. A large table showing these demographic characteristics – region of the country, SES, urban-rural --was then produced ostensibly to show that the sample was indeed representative of the nation. Although the entries in this table looked bizarre to me and my colleagues, they did not look that bad to someone unfamiliar with large-scale testing. For example, there were now more high SES Hispanics and Blacks than high-SES Whites. This is not the world in which we live.

But here is where it gets really interesting. Suppose demographic variables such as SES were indeed distributed in this way. Would that produce coincident score distributions? Surely Dr. Naglieri must believe this to be the case or he would not be making these claims. I asked one of my graduate students to carry out a full simulation of this procedure. We assumed a mean of 90 for the low scoring group and 100 for the high scoring group (SD 16). We then used social class as the demographic variable, made the distributions of social class conform to those given by the census bureau, and then made the correlation between SES and our test score equal .3 (which is population value). Next test scores for the two groups were re-weighted to be coincident (new mean 95) and then tallied the new distributions of SES. We then went backwards and said: suppose SES were so distributed? How much reduction would we see in the 10 point score difference? Answer: less than one point. In other words, it doesn't work.

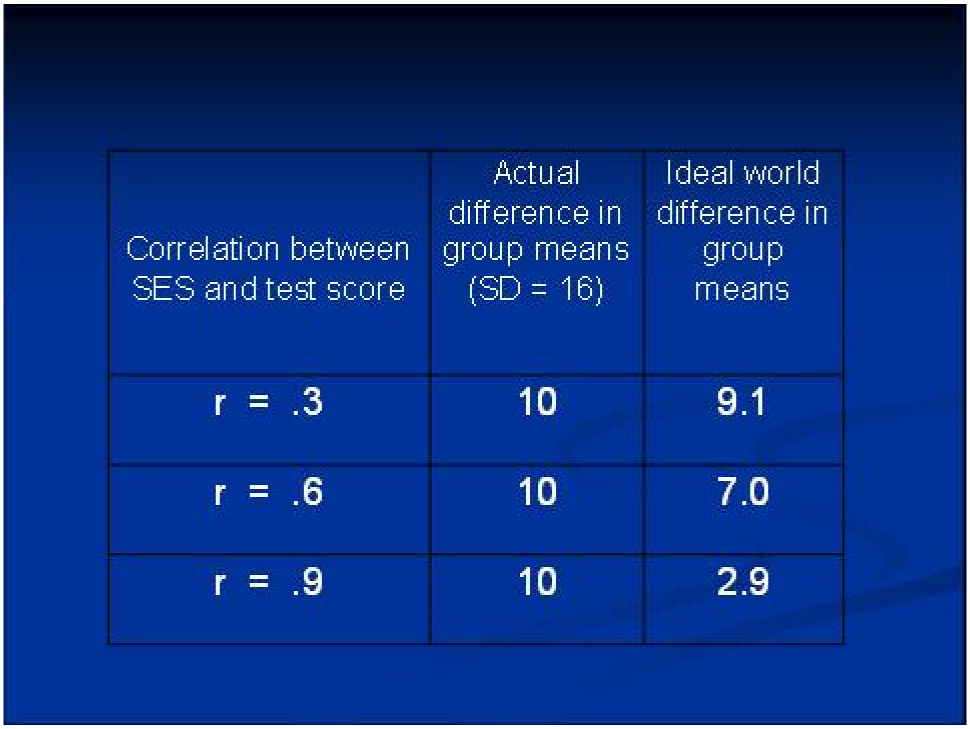
Why? In statistics it's called inverse probabilities. In logic, it's called the fallacy of affirming the consequent. Here is an example.

Adolescents who are convicted of crimes tend to perform poorly in school.

Sue is performing poorly in school.

Therefore, Sue has a criminal record.

The conclusion is most likely to be wrong when variables have low correlations. SES correlates about .3 with scores on ability tests. If the correlation were .6 the 10 point difference would be reduced to 7.2 points. If the correlation were .9, then it would reduce to 3.1 points.



Correlation between SES and test score	Actual difference in group means (SD = 16)	Ideal world difference in group means
r = .3	10	9.1
r = .6	10	7.0
r = .9	10	2.9

I can understand how someone who does not understand much about statistics might fool themselves into thinking that this sort of re-weighting of the data is an interesting thing to do. This is why researchers make explicit what they have done so that others can tell them that their methods are spurious, or at least so that others can replicate their results. But readers were not told that the data had been altered in this way – either in the original article or in the reply to my critique of that article. Most astonishingly to me, I still get emails from people who have attended recent presentations in which these same artificial data have been presented as if they were real. One of the less pleasant duties I had this year was to sit on a committee that was charged with the difficult task of deciding whether a researcher had committed academic fraud. In the process, I learned the legal definition of fraud: Fraud is said to occur when one knowing presents information knowing that others might reasonably misinterpret it.

Enough of this. My initial assertion is the conclusion here – your task has been made much more difficult than it should be because of misleading claims made for nonverbal tests. Nonverbal test do have a role to play in the process of identifying academically talented students. But they should never be used as the primary screening measure. Height and weight are correlated. You can predict weight from height but only with much error. If you want to know how much children weigh, then weigh them if you can. It is not fairer to measure height for all children just because you cannot weigh some of them. If you find all of this is somewhat confusing, then you might find the brief summary in the NRC monograph helpful.

The Process

I have argued that the best way to identify students who are likely to excel in particular domains is to measure the aptitude variables that best predict subsequent accomplishments in that domain. I would like briefly to show you how this can be done. The procedure requires that one know how to use a spread sheet. Detailed directions are provided in the sample data set on my website. Here are the steps for using only one variable.

1. Get the data that you need into the spreadsheet.

   Student name or ID

   Opportunity to learn index (OTL)

   National PR or other norm-reference test scores

2. Sort the data by PR (to get local ranks)

3. Sort again by OTL and then PR (to get rank within OTL)

If using more than one variable (e.g., reasoning abilities and achievement in a domain)

1. Get the data into the spreadsheet, including scaled scores.

2. If scaled scores are from different tests, then put them on the same scale. This can be done by converting them to z scores using the “standardize” function on Excel.

3. Combine these z scores – usually by summing

4. Sort on the basis of this composite score

For a more detailed explanation and examples, see the data set on my web site. For examples that include teacher ratings, see the more recent paper with Joni Lakin.

Implications of an Aptitude Perspective

One of the most important benefits of an aptitude perspective is that it encourages one to focus on the development of expertise rather than the possession of an innate attribute. In this, I follow many others in the field who have argued for the importance of such a perspective. I particularly like the balanced approach that Donna Mathews and Joanne Foster have taken on this. Unlike some who advocate an expertise model, they clearly recognize the importance of cognitive abilities and other aptitudes. A focus on expertise leads one focus on education over the long haul. Where are we going? How can we help children get there? It also encourages us to think about giftedness – especially in young children – less as a permanent state of being and more in terms of the status of the child's current development of some of those abilities, knowledge, and other predispositions that are needed either for the attainment of expertise or that directly reflect its development.

Howard Gardner's work is probably best understood in this way. His intelligences are really different varieties (or aspects) of expertise that are valued by society. All require multiple aptitudes for their development. Further, a developmental perspective helps us understand that we can only see a short distance down the path. The path that leads from beginner to expert has multiple phases and junctures, each of which often demands that the learner bring to bear new aptitudes. For example, expertise in music usually begins with evidence of the ability to retain and reproduce sequences of tonal patterns and rhythms. But then the child must master complex physical skills required to express those patterns – by for example, in playing the piano or the violin. However, some who master both of these phases falter when they must learn to read music. Those who succeed here must then understand music theory. And finally, composition requires a whole new set of abilities. A child may excel at any point along this sequence, yet falter at the next. More importantly, some who falter early excel at later steps.

Indeed, correlation studies of the development of all abilities show this. This is why it is critical continually to reassess students' abilities and competencies. As they grow, some move up, other move down. Longitudinal studies that follow only those who excel at one point – especially an early point – capture only a part of the population. Humphrey's estimated that only about a third of the children who true ability test scores fall in the top 3% of the distribution at age 7 will still be there at age 17. Of course, we never have true scores. Error-encumbered observed scores show even more regression.

Probably the single greatest need in the field is for longitudinal research that follows all children, not just those who are identified as gifted at one point in time. These studies need to measure more than general ability and basic skills on school achievement tests. They should also track motivation, persistence, interests, and other aptitudes that together seem to be required for development of high levels of competence in different domains. Evolutionary biologists are fond of pointing out that one gets a very different picture of the evolution of humans by following backwards the twig on the evolutionary tree that we occupy than by beginning at the other end and attempting to locate that twig. What in retrospect looks like a straight line is actually a very complex system with many decision points.

Suggestions for Policy

1. What are the purposes of the TAG program?

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

2. *What are the proper norm groups to use when making inferences about aptitude? About achievement?*

One needs common national and local standards for the measurement of current achievement and within-group standards for the measurement of aptitude (where "group" is defined by opportunity to learn).

3. *What educational treatment options are available?*

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

4. *Obtain the most reliable and valid measures of achievement, reasoning abilities, and other aptitude variables for all students.*

Less reliable tests will always show smaller group differences than more reliable tests (especially when reliability is estimated by consistency across forms and occasions rather than by an internal consistency coefficient). Less valid tests will often (but not always) show smaller group differences as well (e.g., nonverbal tests). Whenever possible, measure the behavior of interest rather than something that merely predicts that behavior. If the goal is to identify students who have unusual talent for particular academic domains, obtain measures of domain-specific accomplishments to date in that domain, the student's ability to reason in the symbol systems required for new learning in that field of study, interest in the domain, and persistence under similar instructional conditions. Keep in mind that aptitude can only be estimated when a student's performance on a task is compared with the performance of other students who have had similar learning opportunities. Common cut scores on less valid and reliable tests may identify significant numbers of minority students, but many of them are not the students who have the greatest academic talent.

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

On norm-referenced tests, examine local percentile ranks for particular domains such as mathematics or science rather than national percentile ranks for composite scores. This will facilitate the goal of providing challenging instruction for all students. When making instructional placements, use local norms to determine the appropriateness of the match. For example, if a student will be placed with seventh graders for mathematics, compare her performance on a test with seventh grade mathematics content to the performance of students in the prospective seventh grade class.

6. *Emphasize that true academic giftedness is evidenced by accomplishment.*

The most unambiguous evidence of giftedness in a domain is stellar performance in that domain. It is not stellar performance on measures that predict performance in that domain. Predictions that one might someday exhibit excellence in a domain are flattering but unhelpful if they do not translate into purposeful striving toward the goal of academic excellence.

7. *Professional judgment is required.*

Just as a curriculum cannot be made teacher-proof, there is no foolproof way to identify those children who will develop the highest levels of academic excellence in adolescence or the highest levels of professional expertise as adults. Simple schemes that establish an arbitrary cut score on an IQ or achievement test are administratively convenient but identify only a fraction of those who will later attain excellence. Identification cannot be made automatic or algorithmic. It will always require good judgment. One of the goals of my work is to assist the next generation of counselors, psychologists, and program coordinators to exercise this judgment more responsibly.