Psychology Science Quarterly, Volume 50, 2008 (2), pp. 269-282

The commonality of extreme discrepancies in the ability profiles of academically gifted students

DAVID F. LOHMAN[1], JAMES GAMBRELL & JONI LAKIN

Abstract

Extreme discrepancies in abilities are more common among the most and least able students than among average ability children. Therefore, procedures for identifying gifted children that deliberately or inadvertently rely on a composite score that averages across ability domains will exclude many children who reason exceptionally well in particular symbol systems. In this article, we first discuss general issues in the measurement of ability profiles. We then introduce a method for categorizing score profiles and finally document the reliability and stability of score profiles using the 2000 standardization data of the Cognitive Abilities Test (Lohman & Hagen, 2001a).

Key words: ability testing, profiles, identification of gifted

[1] David F. Lohmann, 224 Lindquist Center, The University of Iowa, Iowa City, IA 52240, USA; email: david-lohman@uiowa.edu

Psychology is ambivalent about score profiles on ability tests. On the one hand, there is overwhelming evidence that human abilities are multidimensional, not unidimensional (Carroll, 1993; Gustafsson & Undheim, 1996). Accordingly, theorists and practitioners have long advocated for ability tests that would measure the cognitive strengths and weaknesses of examinees. On the other hand, empirical evidence on the reliability and utility of profiles is at best weak, especially for subtest scores on individually administered ability tests (Watkins, Gluting, & Youngstrom, 2005).

Gifted education is similarly ambivalent about profiles. Although many have advocated for multidimensional theories of giftedness (Feldhusen & Jarwan, 2000; Gagné, 2003; Sternberg, 2003), children are often admitted to programs for the gifted and talented using measures of general ability (Assouline, 2003). Some programs require an IQ score of at least 130. Other programs follow the recommended practice of collecting multiple sources of information but unwittingly collapse it into a measure of general ability. For example, scores from different tests, classroom grades, and teacher ratings are collected. Points are assigned to each (often using somewhat arbitrary rules) and then summed. Admission is based on the total number of points, which estimates the general factor common to all of the measures.

On the other hand, some programs admit students who display excellence in any one of several different cognitive or academic domains (Tannenbaum, 2003). However, unless instruction is appropriately differentiated for such students, these selection policies treat each test or test battery as interchangeable measures of a common ability construct. This ignores the often substantial specific variance in test scores for different batteries. For example, the scores on nonverbal tests may be accepted in lieu of scores on more verbally-loaded tests for examinees with limited fluency in the language of the test. However, absent information on the verbal or quantitative reasoning abilities of the student, many students selected on only the nonverbal test score will have little likelihood of succeeding in an academically challenging program (Lohman, 2005b; Lohman & Lakin, 2007).

Treating different tests as exchangeable measures of general ability is related to the widespread belief that many gifted children display essentially high, flat ability score profiles. However, Achter, Benbow, and Lubinski (1997) argued that, when reliable tests with adequate ceilings are used, most gifted students exhibited significant strengths and weaknesses in abilities that had important long-term consequences.[2] For example, profiles on ability tests administered at age 13 predicted the undergraduate majors students chose and the advanced degrees that they obtained (Lubinski, Webb, Morelock, & Benbow, 2001; Park, Lubinski, & Benbow, 2007). Thus, ability profiles have demonstrated utility for both research and guidance counseling with gifted students.

In this paper, we show that not only can ability profiles be measured reliably; but that extremely large differences between verbal, quantitative, and figural reasoning abilities are much more common among the most able children than among average ability children. Failure to attend to the profile of students' reasoning abilities and to rely instead only on composite scores or other measures of general ability excludes those academically talented students whose relative weakness in one domain reduces their composite score below the established cutoff for admission. But first, we discuss general issues in the measurement of profiles and how a classification scheme can increase their educational utility.

[2] Even if a test has an adequate ceiling, using percentile ranks will mask differences that scale scores would reveal.

Unavoidable tradeoffs in profiles

Understanding how one might best capture information in score profiles requires understanding of the statistical properties of profiles and the tradeoffs that invariably attend efforts to increase the dependability and utility of that information.

Whenever correlations among tests are positive, the general factor common to all tests captures the most variation in the correlation matrix. The higher the correlations among the tests, the larger the general factor will be. It has been argued that this general factor variance must be controlled in order to understand the unique information in score profiles (Gustafsson & Snow, 1997). Once the general factor is removed, the remaining test-specific information comes in the form of a difference score for each test. Difference scores are generally much less reliable than the scores from which they are derived, especially when those scores are highly correlated and their variances are constrained to a common value (Feldt & Brennan, 1989). It is this unhappy reality that renders suspect much of the observed score variation reflected in profiles.

There are three ways to increase the information in the profile: (a) decrease noise in the profile of observed scores by estimating the profile for universe scores or latent variables; (b) reduce the impact of the general factor by reducing the correlations among subtest scores; or (c) increase the reliability of subtest scores. We consider each of these strategies.

Latent score profiles. The estimation of universe scores (Cronbach, Rajaratnam, & Gleser, 1972) would seem an obvious way to reduce the contribution of measurement error and thereby increase the dependability of the profile. The profile of universe scores can differ importantly from the observed score profile when the generalizability (i.e., reliability) coefficients for scores in the profile vary. However, when the generalizability coefficients for observed scores are similar, then the observed scores will regress similarly to their common mean and the profile of universe scores will simply be a flatter version of the profile of observed scores (Cronbach et al.). Multivariate methods for estimating score generalizability augment estimates of latent scores for each variable with information reflected in its covariances with other scores in the profile. However, the improvement is unlikely to be practically significant when the univariate generalizability coefficient for the variable is large and/or its correlations with other scores in the profile are small (Cronbach et al.).

The estimation of latent variables adds an additional step. Typically, the number of latent variables is less than the number of observed scores. Therefore, in addition to controlling for errors of measurement, latent scores also reduce the dimensionality of the score space. Reducing the number of scores in the profile generally enhances the interpretability of the profile and increases likelihood of replicating it (Flanagan & Ortiz, 2001; Gustafsson & Snow, 1997). In the context of ability testing, this means that profiles of subtest scores are rarely as replicable or meaningful profiles of weighted composites of the subtests.

Reducing g. Although measurement experts have emphasized techniques for reducing the impact of measurement error through the estimation of universe or latent scores, developers of ability tests have also attempted to improve the dependability of score profiles by reducing the correlations among subtests in the battery. Although a diverse, moderately correlated collection of subtests better estimates the broad group factor common to all tests, reducing the correlations among subtests generally results in subtests that have lower correlations with external criteria that are not narrowly related to the abilities measured by the subtest. For example, less simple, speeded spatial tests show lower *g* loadings than more complex spatial tests that define the General Visualization (Gv) factor in Carroll's (1993) theory. However, simple spatial tests have never shown the external validity that the more complex spatial tests show, either in prediction studies (McGee, 1979) or in aptitude-by-treatment interaction studies (Cronbach & Snow, 1977). The paradox, then, is that although reducing the common variance among tests makes the profile more dependable, it can also render the scores themselves less useful.

Increase test reliability. The third strategy for increasing the dependability of profiles is to increase the proportion of specific but reliable variance in the separate test scores. Measurement error is minimized by lengthening each test and by sampling more thoroughly from the universe of tasks that elicit the abilities the test purports to measure. This is difficult to do when the test battery must be administered in a relatively short time and, simultaneously, measure the ever growing list of abilities that characterize modern theories of intelligence (Frazier & Youngstrom, 2007). Further, the relationship between test length and reliability is nonlinear, and so meaningful increments in reliability require adding an increasingly large number of items. However, if the test battery is focused on the measurement of two or three abilities, then meaningful increases in subtest reliability are possible.

The Cognitive Abilities Test

Thorndike and Hagen (1971; 1984; 1992) followed this last strategy when developing the Cognitive Abilities Test (CogAT). Rather than measure a broad range of different abilities, they focused on those reasoning abilities with predictive validity for educational achievement: verbal reasoning, quantitative reasoning, and nonverbal/figural reasoning. This split foreshadowed Carroll's (1993) conclusion some twenty years later about the nature of fluid reasoning ability. Carroll's three-stratum theory posits a large array of specific, or stratum I, abilities. These narrow abilities may be grouped into eight broad, or stratum II, abilities. Stratum II abilities in turn define a general (*g*) cognitive ability factor at the third level. Importantly, the broad abilities at Stratum II vary in their proximity to the *g* factor at stratum III. The factor closest to *g* is the broad fluid reasoning or Gf factor. Carroll's analyses of the fluid reasoning factor show that it in turn is defined by three reasoning abilities: (1) sequential reasoning – verbal, logical, or deductive reasoning; (2) quantitative reasoning – inductive or deductive reasoning with quantitative concepts; and (3) inductive reasoning – the core component of most figural reasoning tasks. These correspond with the three CogAT batteries: verbal reasoning, quantitative reasoning, and figural/nonverbal reasoning.

Administering all three batteries of CogAT requires 90 minutes of testing, in addition to time for directions and practice problems for the three tests in each test battery. If given two hours to test students' abilities, most psychologists would not administer a battery of nine different reasoning tests. Instead, they would try to represent a much broader slice of the Stratum II or Stratum III abilities in Carroll's model. Because of this, most psychologists would not have time to obtain reliable measures for the distinguishably different abilities to reason with words (and the concepts they can signify), with numbers or symbols (and the concepts they can signify), and stylized spatial figures (and the concepts they can signify). Instead, most would measure either a composite reasoning factor or only one of the three subfactors of Gf – typically, inductive reasoning with figural stimuli.

To improve the measurement of the three reasoning factors, each of the three CogAT batteries contains three subtests that follow different item formats. The three subtests in each battery are jointly scaled using a Rasch (1960/1980) model, and a total score is computed. As shown in Table 1, median correlations among the total scores on the three batteries range from *r* = .70 to *r* = .76, indicating a strong general factor (Lohman & Hagen, 2002). However, the reliabilities of the batteries are considerably higher: median KR-20 reliability coefficients range from $r_{xx'}$ = .93 to .95 and median parallel forms coefficients for tests administered on different days range from $r_{xx'}$ = .90 to .91. This means that from one-third to one-half of the reliable variance on each battery is not shared with one of the other two test batteries. Profiles attempt to capture this non-shared but reliable variance.

Table 1:
Median Correlations (Across Test Levels A – H) between Batteries and, on the Diagonal, KR-20 (Parallel Forms) Reliabilities for Form 6 of CogAT

	Verbal	Quantitative	Nonverbal
Verbal	.95 (.91)	.72	.70
Quantitative		.93 (.90)	.76
Nonverbal			.95 (.90)

Information in score profiles

Every score profile contains three kinds of information: altitude, scatter, and shape (Cronbach & Gleser, 1953). Altitude refers to the overall height or level of the score profile. It reflects the influence of the general factor common to the scores in the profile. If all tests contribute equally to that general factor, then altitude can be indexed by the average of the separate scores. Scatter refers to the variability of subtest scores. A common index of scatter is the standard deviation of subtest scores about the examinee's mean score across those subtests. Finally shape refers to the particular pattern of elevations and depressions in the profile. Profiles with the same amount of scatter can have quite different shapes.

Altitude. There is little controversy about the measurement of altitude. It is best estimated by a weighted sum of subtest scores where the weights are proportional to the loading of each test on the general factor. More commonly, it is estimated by the sum (or centroid) of raw or scale scores in the profile.

Scatter. Clinicians have long speculated on the potential diagnostic utility of scatter among the test scores on a battery. However, empirical analyses have met with little success (Watkins et al., 2005). Part of the problem is that scatter masks differences in profile shape. For example, students with different types of disabilities may have profiles that differ in shape but have similar scatter. But the larger problem may be that a profile of seemingly similar scores may be only slightly more probable than a profile with more obvious peaks and valleys. What is needed, then, is an estimate of the extent to which the scatter of scores is unusually large. Unusual scatter means unusually large within-person variability of subtest scores. The expected within-person variance of subtest scores is easily obtained by averaging the within-person variances of subtest scores. This expected variability of profiles scores can also be estimated directly from the basic variance-covariance matrix of subtest scores (Brennan, 2001). Confidence intervals for the expected profile variance can then be computed using the χ^2 statistic. If the typical within-person variability is large, then composite scores capture only a portion of the information in the test scores. Whether this additional information is dependable depends on the reliability of the subtests. Whether it is useful also depends on the size of the differences between subtest scores.

Shape. The shape of the profile can be indexed by the similarity of the score profile with some standard. This standard can be defined empirically or logically. An empirical standard could be established by Q factor analysis or by any other procedure that identifies the dimensionality of the profile space (such as the PAMS analyses of Kim, Frisby, & Davison, 2004). Then, the similarity between the profiles for individuals and each of the latent or characteristic profiles is estimated and individuals are classified according to the latent profile that differs least from their observed profile.[3] A logical standard can be established by enumerating the patterns that could be observed. This is a reasonable option only when the number of test scores in the profile is relatively small. For example, for the three CogAT scores (V, Q, and N) there are three profiles that show a strength on one battery (V+, Q+, or N+), three that show a weakness (V-, Q-, or N-), and six more that show both a strength and a weakness (V+Q-, V+N-, Q+N-, Q+V-, N+V-, and N+Q-). All examinees that show differences across subtests can be classified in one of these 12 categories. However, such a classification system would become unwieldy with a larger number of subtests.

The CogAT6 profile system

Assuming that one has constructed a battery of tests that reliably measure multiple dimensions, how can the score profile be represented? A system for representing profiles must capture all three sources of information (altitude, scatter, and shape). Further, it must do so in a way that is useful to test users. Inevitably, communication requires categorization not only of the profiles into empirically or logically derived categories, but also of the similarity of the scores for a given individual to the profiles. Users want to know, for example, if the child displays an educationally important pattern of strengths or weakness across the domains reflected in the score profile. Different patterns of strengths and weaknesses sometimes have well-established educational implications (e.g., much stronger verbal than spatial abilities). However, these educational implications can differ for low altitude (i.e., low ability) profiles and high altitude (i.e., high ability) profiles. Further, the implications might not be the same when the differences between scores are unusually large (i.e., show more scatter).

The system that we developed to index profiles on CogAT attends to all three types of information in profiles: altitude, scatter, and pattern. Our goal was also to find a way of encapsulating this information that would enable test users to get quickly from the profile to sug-

[3] A PAMS analysis of the nine CogAT subtests would show two major dimensions: (a) a contrast between the three verbal and the three nonverbal subtests and (b) a contrast between the three quantitative subtests and the other six subtests. This is the same pattern that would be observed in the first two unrotated factors of a factor analysis of the correlations among the nine tests.

gestions for interpreting and using the information to inform instruction.[4] This required that we categorize profiles in ways that would, on the one hand, well summarize the information in them while, on the other hand, not overwhelm the user with an infinite number of variations.

Score profiles are summarized in a simple code. Example profiles are 3A, 9B(V-), and 6C(V+Q-). The number is the student's median age stanine on the three batteries. Stanines range from 1 (lowest 4 % of scores in the distribution) to 9 (highest 4 % of scores in the distribution). The median stanine estimates the overall level or altitude of the profile. We chose the median over the mean because it better represents central tendency when one of the scores differs markedly from the other two. This is especially important at the extremes of the distribution.

Pattern was indexed next. The first letter tells whether all three scores were at the sAme level (an "A" profile), whether one score was aBove or Below the other two scores (a "B" profile), or whether two scores showed a significant Contrast (a "C") profile. In the examples above, 3A means that the median age stanine was 3 and that the three scores did not differ significantly from one another. The second example, 9B(V-), means that the median age stanine was 9 and that the score on the Verbal Battery was significantly lower than the scores on the Quantitative and Nonverbal batteries, which did not differ from each other. The last profile, 6C(V+Q-), shows a relative strength on the Verbal Battery and relative weakness on the Quantitative Battery. The median stanine was 6 – slightly above average.

In constructing profiles, scores were considered significantly different only if the difference met two criteria. First, the 95 % confidence intervals for the two scores could not overlap. These intervals were based on the estimated IRT conditional standard error of measurement (SEM). Since IRT SEM is larger for scores near the extremes of the distribution, this means that observed differences had to be larger for the most (and least) able students in order to be considered significantly different. The average standard errors of the difference between batteries (across all score levels) were 5.5, 5.2, and 5.4 (for Verbal versus Quantitative, Verbal versus Nonverbal, and Quantitative versus Nonverbal, respectively). Second, the difference had to be greater than 10 points on the SAS scale ($\overline{X}$ = 100, s = 16). This insures that small differences near the mean of the distribution were considered practically insignificant, even though they might be statistically significant.

Finally, the degree of scatter in the profile was indexed by re-labeling profiles with extremely large differences between scores as Extreme (E) profiles. These occur when two scores differed more than 24 points (or 1.5 *s*) on the SAS scale. For example, 8E (N-) means that the median stanine was 8 and that the score on the Nonverbal Battery was at least 24 SAS points lower than the score on one of the other two batteries.

Frequency of Occurrence of Different Score Profiles

Are some profiles more common than other profiles for the most or least able students? Table 2 shows the percentage of students in the 2000 CogAT standardization sample who had different score profiles on the CogAT multilevel battery. The first column shows the

[4] Instructional suggestions are provided in the Short Guide for Teachers and in an interactive profile interpretation system. Both can be accessed at www.cogat.com.

percentages for all 115,133 students in the standardization sample. If all three CogAT reasoning scores measure a single ability, then the majority of students should have approximately equal scores on the Verbal, Quantitative, and Nonverbal batteries. Here, this would be represented by an "A" profile. However, only 39.8 % of the entire student sample had a flat or "A" profile. Stated the other way, the majority of students showed significantly uneven profiles in reasoning abilities. Of this majority, 35.8 % showed a significant, but not extreme, strength or weakness in one area. (See the "Total B" row.) Another 7.7 % showed an extreme strength or weakness. (See the "Total Extreme B" row.) Finally, 16.7 % showed a significant (13.4 %) or extreme (3.3 %) contrast between two scores ("C" and "Extreme C" rows).

The second column of Table 2 shows the corresponding percentages for the 3,062 students who obtained a stanine score of 9 on at least two of the three CogAT batteries. This addresses the question of whether some profiles are more common among gifted students than among the general student population. Once again, only 39.2 % showed a flat or "A" profile, in spite of the fact that the group contained only those who obtained a stanine score of 9 on two of the three batteries. However, 36.6 % showed a significant (21.4 %) or extreme (15.2 %) weakness on the third battery. Put differently, the most able students were about as likely to show a significant or extreme weakness on the third battery as they were to show a score similar to the score they obtained on the other two batteries.

Figure 1 provides another perspective on the relative frequencies of different profiles. Flat ("A") profiles were slightly more common among the least able students (median stanines of 1 or 2) but constant thereafter at about 40 % of the cases. A significant or extreme strength [B+ or E(B+)] was most common among students with a median stanine score of 1 and least common for students with a median stanine score of 9. A significant or ex-

Table 2:

Percent of All Students and of High-Scoring Students (Median Stanine = 9) Obtaining Different Profiles of Verbal, Quantitative, and Nonverbal Reasoning Abilities on the CogAT Form 6 Multilevel Battery

Profile	All Students (*n* = 115,133)	Median Stanine of 9 (*n* = 3,062)
A	39.8	39.2
B	35.8	27.9
V+,Q+, N+	(18.3)	(6.5)
V-, Q-, N-	(17.5)	(21.4)
C (All)	13.4	11.0
Extreme B	7.7	18.3
V+, Q+, N+	(4.3)	(3.1)
V-, Q-, N-	(3.4)	(15.2)
Extreme C (All)	3.3	3.6
Total	100.0	100.0

Note. V = Verbal; Q = Quantitative; N = Nonverbal; A = All three scores at approximately the sAme level; B = One score aBove or Below the other two scores; C = Two scores Contrast significantly; Extreme = Scores differ by at least 24 SAS points.

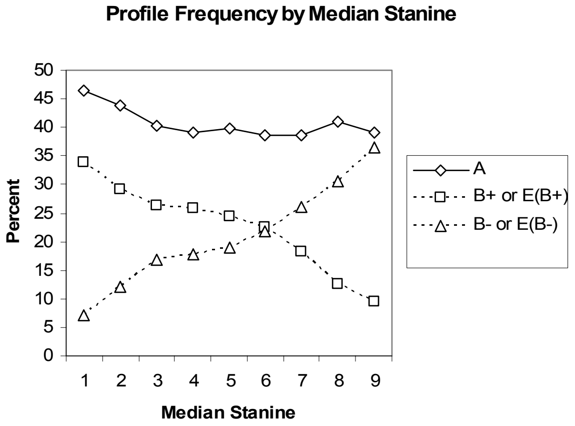

Figure 1:
Profile frequency by median stanine. Flat or "A" profiles (diamonds); a significant strength (B+) or extreme strength [E(B+)] (squares); a significant weakness (B-) or extreme weakness [E(B-)] (triangles)

treme weakness [B- or E(B-)] shows the opposite pattern. The most able students were much more likely to show this profile than other students. A moment's reflection on the nature of correlation will show that this must be the case. For low-scoring students, an even lower score on the third battery is much less likely than a higher score. Conversely, high scoring students are much more likely to show a large weakness than to show a strength on the third battery. This is what is meant – but not explained – by "regression to the mean."

Regression to the mean will be observed whenever the correlation between two variables is less than 1.0. Some of this regression reflects error of measurement. Here, the reliabilities of the three CogAT scores are much higher than their correlations with each other (see Table 1). Therefore, the largest contributor to regression is not error of measurement but the fact that each of the three test batteries measures somewhat different abilities. This means that the probability of obtaining an extremely high score on all three test batteries is much less likely than the probability of obtaining two high scores and a lower score on the third battery.

The greater frequency of one extremely low score for high-altitude profiles is not unique to CogAT. Indeed, the lower the correlation among scores in the profile (for reasons including multidimensionality as well as measurement error), the more common it will be. This is one of the main reasons why the CogAT authors have always admonished test users not to use the three-battery composite score to screen children for admission to programs for the gifted (Lohman & Hagen, 2001b; Thorndike & Hagen, 1984; 1992). Even very capable students are likely to have one battery score that is low enough to bring down their averaged composite score. This is particularly problematic when the low score is on the Nonverbal Battery, which is least related to success in verbally demanding academic domains.[5] Rather than using only an estimate of *g*, the better procedure is to match the selection criteria with the demands of the educational program (Lohman, 2005a; Renzulli, 2005).

[5] Indeed, once the g variance has been accounted for, the Nonverbal score sometimes enters with a negative weight in the prediction of academic achievement (Lohman, 2005b).

The reliability of score profiles

Although two scores may differ significantly on the day that students were tested, would the same differences be observed on another form of the test that was administered one to three weeks later? Since profiles are composed of difference scores, we can estimate the stability of the profiles of all three scores from the stability of differences between pairs of scores on two of the three batteries. If the difference between two scores is not dependable, then a profile of multiple differences cannot be dependable. Since each B profile contains two significant differences, the probability that both differences will be significantly greater than zero is less than the probability that either one of them will differ significantly different from zero. For example, if we establish a 90 % confidence interval, then the probability that V>Q and V>N would be the product of the separate probabilities (i.e., .10 x .10 = .011) if the two were independent events. Both comparisons contain V and so these are not independent events and so the probability is somewhat higher.

We estimated the stability of each pair of difference scores among the three batteries using the standard formula for the reliability of difference scores (Feldt & Brennan, 1989, p.118). Reliabilities of separate scores were estimated from the correlations between parallel forms of CogAT administered on separate occasions. We then computed the standard error of each of the three difference scores using the observed standard deviation of that difference score. Finally, we computed confidence intervals corresponding to 1.0 probable error (50 % confidence interval), 1.28 SEM (80 % confidence interval) and 1.645 SEM (90 % confidence interval). These are reported in Table 3.

Even though the reliabilities of the difference scores across forms and occasions (median r = .636) are considerably lower than the between-forms reliabilities of the separate scores on three batteries (median r = .90), the 90 % confidence intervals for the difference are not much larger than the confidence intervals for inferring a significant difference between scores at Time 1. This is because the variability of difference scores reported in Table 3 (median 11.5) is considerably less than the variability of the separate battery scores (fixed at 16). The table shows that differences of 14 (or more) points on the SAS scale ($\overline{x}$ = 100, s = 16) obtained on Form 6 of CogAT are quite likely (p < .05) to be observed if Form 5 were to be administered within a few weeks. Even differences as small as 10 SAS points (the minimum difference required for declaring differences significant at Time 1) would likely still be considered significantly different at retest with an 85 % confidence interval.

Although these analyses usefully inform the replicability of differences between scores on particular batteries, they do not tell us how stable the profile classifications might be. We

Table 3:
Reliabilities (Across Forms and Occasions), Standard Deviations, and Various Standard Errors for Difference Scores between the Three CogAT Scores

Difference Score	$r_{DD'}$	s_D	SEM_D	90 % CI	95 % CI
Verbal - Nonverbal	.672	12.38	7.1	11.7	13.9
Verbal - Quant	.636	11.53	7.0	11.4	13.6
Quant - Nonverbal	.574	10.72	7.0	11.5	13.7

would not expect as much consistency for a classification scheme with 13 categories as for a scheme with only one score per examinee (i.e., a measure of profile altitude or *g*). By way of comparison, it is useful to consider the retest consistency of a single test score. Assume a correlation of $r = .80$ across occasions. The probability that an individual below the median on Occasion 1 will be below the median at Occasion 2 is about .79. But this is a very lenient test. Suppose we ask that the retest score be within ten percentile points of the original test. Now the probability drops to a median value of .17. Thus, the fact that profile classifications, which require replication of multiple scores within narrow boundaries, show lower consistencies across occasions than composite scores is easily misinterpreted. Even highly correlated scores show less consistency than most people expect.[6]

Another way to view profile consistency is to consider various degrees of replication of the profile. For example, profiles B(V+), C(V+Q-), C(V+N-), and the three extreme versions of these profiles all show a verbal strength. One could examine not only exact replication of a profile (e.g., B(V+), but also replication within the family of similar profiles (any one of the six profiles that contain V+). Replicability of profiles for empirically grounded profile schemes could also be judged by distances to families of similar prototypical profiles.

Finally, an important form of profile replication (or generalization) can be obtained by comparing the profile of ability test scores with measures of achievement in the corresponding domains. This form of cross-battery assessment (Flanagan & Ortiz, 2001) is readily available to many educators. The observation of an extreme strength or weakness in quantitative reasoning that does not also appear on measures of mathematics achievement should be suspect. This is one of the reasons why the identification of academic talent is better made by the judicious combination of measures of ability and achievement than from either alone (Lohman & Korb, 2006; Lohman & Renzulli, 2007).

Summary

We had several goals in this paper. The primary goal was to show that extreme discrepancies in abilities are much more common among the most (and least) able students than among average ability children. Therefore, procedures for identifying academically talented students that either deliberately or inadvertently rely on a single composite score that averages across ability domains will exclude many children who reason well in particular symbol systems. Even students with strong ability to reason in two symbol systems can have scores in the third area that bring down their composite score. Consistently high scores across multiple domains is not a necessary feature of giftedness. True, those who exhibit high scores in all domains tested are very able. But they are not the only gifted students who warrant special attention.

Our second goal was to discuss general issues in the measurement of ability profiles. Profiles differ in altitude, scatter, and shape. Methods for classifying profiles must capture all three dimensions. Reliability of the profile can be increased by estimating a profile for universe scores or latent variables, decreasing the correlations among the scores in the profile,

[6] These probabilities are readily obtained from standard tables of prediction efficiencies for correlations. See, e.g., “Tables of prediction efficiencies” at
http://faculty.education.uiowa.edu/dlohman/pdf/tables_of_prediction_efficiencies.pdf.

or increasing the reliability of the separate scores. Although most discussions of profiles have emphasized the first two methods, we demonstrate the value of meaningfully increasing the reliability of the separate scores in the profile.

Our third goal was to introduce a simple method for enumerating and categorizing score profiles. We illustrated this scheme using the 2000 standardization data of the Cognitive Abilities Test (Lohman & Hagen, 2001a). A logically derived scheme works well for tests such as CogAT because (a) there are only three scores to consider, (b) correlations among the three batteries are reasonably uniform, and (c) reliabilities of the three scores are uniform and much higher than their correlations with each other. When we examined the reliability of the difference scores that compose these profiles, we found that two scores that differ by 14 or more points on the SAS scale are quite likely to differ significantly on retest. Finally, we argue that the implicit standard used to judge profiles assumes that total scores show greater consistency in classification accuracy than is observed even when total scores are highly correlated. Nonetheless, even under the best of conditions, profiles are less dependable than total scores across all subtests. Therefore, selection into a particular kind of instruction should not be made on the basis of a single administration of a test battery. Replication with parallel tests or, better yet, generalization to similar measures of ability or achievement should be required as well.

In their defense of profile analysis, Flanagan and Ortiz (2001) argue that many of the problems that have plagued interpretation of ability test profiles can be ameliorated by (a) interpretation within the context of a well-validated theory of abilities, (b) basing the profile on scores that combine two or more subtests rather than on individual subtest scores; (c) incorporating normative information on profile scores, particularly those classified as “weaknesses”; and (b) not expecting stability in scale score profiles over an extended period of time. The analyses reported here incorporate all of these recommendations: the structure of the test mirrors Carroll’s (1993) model of broad fluid reasoning factor; the nine CogAT subtests are first combined into three composite scores; normative information is incorporated by attending explicitly to the altitude of the profiles; and estimates of profile stability are based on one-month retest scores on a parallel form of the test. There is one important difference, however. Flanagan and Ortiz argue that a “weakness” that falls within the normal range of functioning is of questionable utility for the diagnosis of disability. This ignores the emerging literature on twice-exceptional students (Moon & Reis, 2004). Although gifted students with such profiles may not be disabled in the same way that average ability students are disabled, radical discrepancies in abilities – which are more common for gifted students than for average ability students – surely affect both their own self perceptions and others evaluations of their abilities. Procedures for identifying gifted students need to be sensitive to these differences as well.

References

Achter, J. A, Benbow, C. P., & Lubinski, D. (1997). Rethinking multipotentiality among the intellectually gifted: A critical review and recommendations. *Gifted Child Quarterly. 41*, 5-15.

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

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

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

Cronbach, L. J. & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456-473.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J. & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Feldhusen, J. F. & Jarwan, F. A. (2000). Identification of gifted and talented youth for educational programs. In K. A. Heller, F. J. Monks, R. Subotnik, & R. J. Sternberg (Eds.) *International handbook of giftedness and talent* (pp. 271-282). Oxford, UK: Elsevier.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. Linn (ed.) *Educational measurement* (3rd ed., pp.105-146). New York: American Council on Education and Macmillan.

Flanagan, D. P., & Ortiz, S. O. (2001). *Essentials of cross-battery assessment*. New York: Wiley.

Frazier, T. W. & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence, 35,* 169-182.

Gagné, F. (2003). Transforming gifts into talents: The DMGT as a developmental theory. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 60-74). Boston: Allyn & Bacon.

Gustafsson, J. E. & Snow, R. E. (1997). Ability profiles. In R. F. Dillon (Ed.) *Handbook on testing* (pp. 107–135). Westport, CT: Greenwood Press.

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

Kim, S.-K., Frisby, C. L., & Davison, M. L. (2004). Estimating cognitive profiles using profile analysis via multidimensional scaling (PAMS). *Journal Multivariate Behavioral Research, 39*, 595-624.

Lohman, D. F. (2005a). An aptitude perspective on talent identification: Implications for the identification of academically gifted minority students. *Journal for the Education of the Gifted*, *28*, 333-359.

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

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

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

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

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

Lohman, D. F. & Lakin, J. (2007). Nonverbal test scores as one component of an identification system: Integrating ability, achievement, and teacher ratings. In J. VanTassel-Baska (Ed.), *Al-*

ternative assessments with gifted and talented students (p. 41-66). Thousand Oaks, CA: Corwin Press.

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

Lubinski, D., Webb, R. M., Morelock, M. J., & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology, 86*, 718-729.

McGee, M. G. (1979). Human spatial abilities: psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin, 86,* 889-918.

Moon, S. M., & Reis, S. M. (2004) Acceleration and twice exceptional students. In N. Colangelo, S. G. Assouline, & M. U. M. Gross (Eds.), *A nation deceived: How schools hold back America's brightest students* (Vol. 2, pp. 109-119). Iowa City, IA: The Connie Belin & Jaqueline N. Blank International Center for Gifted Education and Talent Development.

Park, G., Lubinski, D., & Benbow, C. P. (2007). Contrasting intellectual patterns predict creativity in the arts and sciences: Tracking intellectually precocious youth over 25 years. *Psychological Science, 18*, 948-952.

Rasch, G. (1980, reprint). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

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.

Sternberg, R. J. (2003). Giftedness according to the theory of successful intelligence. N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed, pp. 88-99). Boston: Allyn & Bacon.

Tannenbaum, A. J. (2003). Nature and nurture of giftedness. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 45-59). Boston: Allyn & Bacon.

Thorndike, R. L. & Hagen, E. P. (1971). *Cognitive Abilities Test (Form 4).* Boston: Houghton Mifflin.

Thorndike, R. L. & Hagen, E. P. (1984). *Cognitive Abilities Test (Form 4).* Chicago: Riverside.

Thorndike, R. L. & Hagen, E. P. (1992). *Cognitive Abilities Test (Form 5).* Chicago: Riverside.

Watkins, M.W., Gluting, J.J., & Youngstrom, E. A. (2005) Issues in subtest profile analysis. In D. P. Flanagan & P. L. Harrison (Eds.) *Contemporary intellectual assessment* (2^{nd} ed., pp. 251-268). New York: The Guilford Press.