Gifted Today but Not Tomorrow? Longitudinal Changes in Ability and Achievement During Elementary School

David F. Lohman and Katrina A. Korb

The term gifted *implies a permanent superiority. However, the majority of children who score in the top few percentiles on ability and achievement tests in 1 grade do not retain their status for more than a year or 2. The tendency of those with high scores on one occasion to obtain somewhat lower scores on a later occasion is one example of regression to the mean. We first summarize some of the basic facts about regression to the mean. We then discuss major causes of regression: errors of measurement, individual differences in growth, changes in the content of the developmental score scale, and changes in the norming population across age or grade cohorts. We then show that year-to-year regression is substantial, even for highly reliable test scores. Different ways of combining achievement and ability test scores to reduce regression effects are illustrated. Implications for selection policies and research on giftedness are also discussed.*

Longitudinal studies of intellectually exceptional students have produced some of the most important findings in the field of gifted education (Lubinski, Webb, Morelock, & Benbow, 2001; Terman & Oden, 1959). However, there is a paradox in the literature on the relationship between estimates of ability in childhood and accomplishments in adulthood. On the one hand, in any group of children, the child who obtains the highest score on a measure of scholastic aptitude is the one who is most likely later to attain the highest level of academic excellence. On the other hand, the student who obtains the highest score is also the person whose test score at some later date is most likely to show the greatest amount of regression to the mean. How is this possible?

Statistically, the paradox of high aptitude being associated both with high accomplishment and large regression effects merely restates what it means for two variables to be imperfectly correlated.

David F. Lohman is Professor of Educational Psychology at the University of Iowa. His primary research interests are the nature and measurement of reasoning abilities, the identification of children with extraordinary academic aptitude, and adapting instruction to the needs and proclivities of learners. Katrina Korb is a doctoral student in Educational Psychology at the University of Iowa. Her primary interest is the measurement of individual differences in cognitive abilities, especially the quantitative reasoning abilities of school children.

Journal for the Education of the Gifted. Vol. 29, No. 4, 2006, pp. 451–484.

The problem, however, is that the sort of probabilistic thinking that is captured in correlations runs counter to the tendency to think categorically about labeled concepts. We speak of *learning-disabled* or *gifted* students as if there were sharp boundaries separating individuals in the categories from those outside of them. Even those who understand that the boundaries are arbitrary often think that if we agreed on the location of the category cut points and had perfectly reliable and valid measures, then category membership would remain constant over time. In the case of academically advanced children, the expectation is that if we could measure giftedness well, then the child who is considered gifted at age 6 would still be considered gifted at 16. If retesting the child at age 8 or 10 suggested a lower score, the typical reaction would be to question either the dependability of scores (especially the latter, lower score) or the validity of the test that produced them. Indeed, it is common practice to administer different ability tests to individuals or groups in the hopes of identifying additional gifted students. The assumption is that any high score is legitimate, whereas lower scores underestimate ability.

Confusions about giftedness thus reflect more than a common fondness for typologies. They also result from assumptions about the nature of intellectual development and the characteristics of tests used to measure that development. For example, the assumption that the child whose performance is unusual at one point in time will be equally unusual at another point in time assumes (a) that errors of measurement do not substantially affect either test score, (b) that growth from Time 1 to Time 2 is constant for all who have the same initial score, (c) that tests measure the same mix of constructs at all points along the score scale that spans the developmental continuum, and (d) that the population of test takers is constant across time. To the extent that these assumptions are not true, then we will see a regression in scores from Time 1 to Time 2.

Regression to the Mean

Regression to the mean occurs whenever scores are not perfectly correlated. The amount of regression in standard scores can easily be

estimated from the correlation between the two sets of scores. The predicted score on Test 2 is simply $\hat{z}_2 = z_1 \times r_{12}$, where $\hat{z}_2$ is the predicted standard score on Test 2, z_1 is the standard score on Test 1, and r_{12} is the correlation between the Tests 1 and 2.

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

The equation for $\hat{z}_2$ can be used to estimate the expected regression in status scores such as IQs. The first step is to convert the IQ to a *z* score by subtracting the mean IQ and dividing by the population *SD* for the test. For example, if the mean is 100 and the *SD* is 15, then an IQ of 130 converts to a *z* score of $\frac{130-100}{15} = 2.0$. If the correlation between scores at Time 1 and Time 2 is $r = .8$, then the expected *z* score at Time 2 is $2.0 \times .8 = 1.6$. This converts to an IQ of $(1.6 \times 15) + 100 = 124$. The expected regression is 6 IQ points. If the IQ were 145, then the expected regression would be 9 IQ points.

The standardized scores used in the equation for $\hat{z}_2$ may be inappropriate if the variance of scores is not the same across occasions.[1] This often occurs when using attainment scores (such as mental age or developmental scale scores) rather than status scores (such as percentile rank or IQ). Whether the variance of attainment scores increases or decreases over time depends on the nature of the abilities that are measured, the dependent measures that are used, and how score scales are constructed. The variance of scores tends to decrease with practice when the domain is closed rather than open (Ackerman, 1989). A closed skill set is one that is relatively small and bounded. For example, learning to count to 10 is a closed skill. Learning mathematics is an open skill. The dependent measure also matters. For example, as individuals learn a new skill, the variance of accuracy scores often declines. However, response speed or other measures of learning and transfer can show improvements with addi-

tional practice. Such scores may show an increase in variance with extended practice.

How tests are scaled can have a substantial impact on whether the variance of scores increases over time. For example, the Iowa Tests of Basic Skills, Form A (ITBS; Hoover, Dunbar, & Frisbie, 2001) and the Cognitive Abilities Test, Form 6 (CogAT; Lohman & Hagen, 2001) were jointly normed on the same national sample. ITBS scaled scores show considerable increase in variance across grades, whereas CogAT scaled scores do not. This is because the ITBS is scaled using a growth model that assumes that individual differences in achievement increase over grades. The CogAT is scaled using the Rasch (1960) model that makes no assumptions about changes in score variance across time. These differences in scaling procedures are masked when status scores such as percentile ranks or standard age scores are reported.

Developmental psychologists recognize that regression to the mean is a pervasive phenomenon when retesting students (Marsh & Hau, 2002; Nesselroade, Stigler, & Baltes, 1980; Phillips, Norris, Osmond, & Maynard, 2002). Regression to the mean is also commonly cited as a problem when working with learning-disabled students (e.g., Milich, Roberts, Loney, & Caputo, 1980). However, this statistical fact of life is less commonly applied to gifted students.[2] Many who recognize the problem often ascribe it entirely to errors of measurement (e.g., Callahan, 1992; Mills & Jackson, 1990). However, measurement error is only part of the picture.

Any factor that reduces the correlation between two sets of scores contributes to regression toward the mean. We discuss five: errors of measurement, conditional errors of measurement, differential growth, changes in the content of the developmental scale, and changes in the norming sample.

Errors of Measurement

Error is the most obvious contributor to regression toward the mean. Sources of error that might lower a score on a particular occasion are called negative error; they include factors such as temporary inattention or distractions when taking the test. Error can also contribute to

higher scores. Examples of positive error are lucky guessing or good fortune in having learned the solutions to particular items. These sorts of seemingly random fluctuations in behavior across situations are what most people understand as errors of measurement.

A larger source of measurement error for most examinees, however, is the particular collection of tasks and items that are presented. For example, the estimate one obtains of a student's reasoning abilities depends on the format of the task (e.g., matrices, analogies, or classification problems) and the particular sample of items presented in each of these tasks. Factor analyses of large test batteries commonly show that the loading of a test on its task-specific factor is often not much smaller than its loading on the factor that it helps define. This means that the scores on the test are as likely to reflect something specific to the task and measurement occasion as something that would be shared with other measures of the same construct. For this reason, measurement experts have long advocated estimating ability using tests that present as many items as possible in many different formats as possible. However, even when tests contain many items in multiple formats, one is almost never interested in the student's score on a particular form of a test that is administered on a particular occasion. The ideal score would be one that is averaged across all acceptable conditions of observation: test formats, samples of items, test occasions, and other conditions of testing.

Several of these factors are varied when scores are obtained for representative samples of students on different individually administered ability tests. Test tasks, test occasions, and perhaps even examiners or other conditions of testing vary. Correlations between individually administered ability tests range from approximately $r = .7$ to .85. For example, Phelps (in McGrew & Woodcock, 2001) reported a correlation of $r = .71$ between the Woodcock-Johnson III General Intellectual Ability score and the Full Scale IQ on the WISC-III for a sample of 150 randomly chosen students from grades 3 to 5. Flanagan, Kranzler, and Keith (in McGrew & Woodcock, 2001) reported a correlation of $r = .70$ between the Woodcock-Johnson III Brief Intellectual Ability score and the Full Scale Score on the Cognitive Assessment System. Roid (2003) reported a correlation of $r = .84$ between the Stanford-Binet V and the WISC-III (see also Daniel, 2000). As shown later, cor-

relations of this magnitude will result in substantial regression when students who receive a high score on one of these tests are administered a different test. For example, given a correlation of $r = .84$, only about half of the students who score in the top 3% of the distribution on one test will also score in the top 3% of the distribution on the other test (see Table 1).

Averaging can reduce the impact errors of measurement. For example, one could compute the average of a student's reading achievement scale scores or ability test scores across 2 years rather than using the score from a single testing. Averaged scores will regress toward the mean, and so the average of two test scores cannot be interpreted using the norms that are derived for a single administration of the test. But, even norms for individual test scores may be misleading. Norms for ability tests—especially nonverbal tests—have changed dramatically over the past 40 years (Flynn, 1987, 1999; Thorndike, 1975). Schools should not use published norms on ability tests that are inadequate (e.g., see Tannenbaum, 1965, on the Culture Fair Intelligence Test) or severely out of date (e.g., the Stanford-Binet L-M). When it is impossible to administer multiple tests of a particular construct, one should endeavor to use tests that present items in multiple formats rather than a single item format. Such tests typically have higher generalizability than those that use a single response format for all items. Finally, as shown later, averaging scores on a domain-specific test of achievement and a test of reasoning abilities in the symbol systems used to communicate new knowledge in that domain can dramatically reduce the amount of regression in test scores.

Conditional Errors of Measurement

Although many researchers understand that the concept of error includes more than random fluctuations across test occasions, fewer understand that the amount of error in test scores is generally not uniform across the score scale. Formulas for estimating the standard error of measurement (*SEM*) from the reliability coefficient generally assume that the variability of errors is constant across score levels. This is a reasonable assumption for most examinees. It is often

Table 1
Proportions of Students Exceeding a Cut Score on Test 1 Who Exceed the Same Cut Score on Test 2, by Cut Score and Correlation Between the Two Tests

	Correlation between tests										
Cut score	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.975
Top 1%	0.13	0.16	0.19	0.22	0.27	0.32	0.38	0.45	0.54	0.67	0.76
Top 2%	0.17	0.20	0.23	0.27	0.31	0.36	0.42	0.49	0.58	0.70	0.79
Top 3%	0.20	0.23	0.26	0.30	0.35	0.40	0.45	0.52	0.60	0.72	0.80
Top 4%	0.22	0.25	0.29	0.33	0.37	0.42	0.48	0.54	0.62	0.73	0.81
Top 5%	0.24	0.28	0.31	0.35	0.39	0.44	0.50	0.56	0.64	0.74	0.82
Top 6%	0.26	0.29	0.33	0.37	0.41	0.46	0.51	0.57	0.65	0.75	0.82
Top 7%	0.28	0.31	0.35	0.38	0.43	0.47	0.53	0.59	0.66	0.76	0.83
Top 8%	0.30	0.33	0.36	0.40	0.44	0.49	0.54	0.60	0.67	0.77	0.83
Top 9%	0.31	0.34	0.38	0.41	0.46	0.50	0.55	0.61	0.68	0.77	0.84
Top 10%	0.32	0.36	0.39	0.43	0.47	0.51	0.56	0.62	0.69	0.78	0.84
Top 15%	0.38	0.42	0.45	0.48	0.52	0.56	0.61	0.66	0.72	0.80	0.86
Top 20%	0.44	0.47	0.50	0.53	0.56	0.60	0.65	0.69	0.75	0.82	0.88

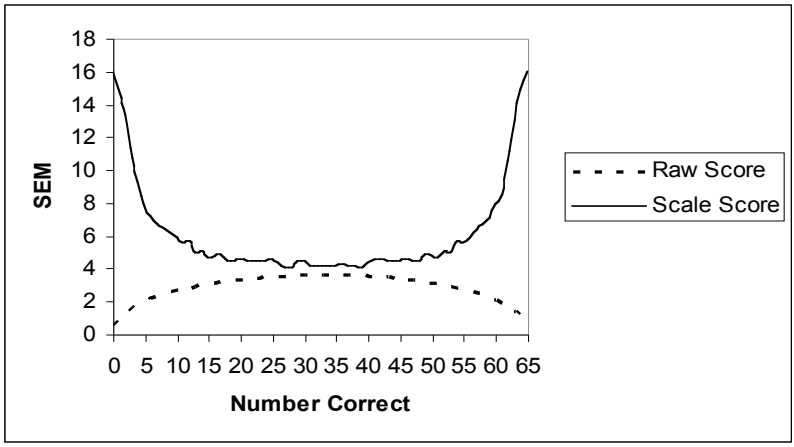

Figure 1. Conditional Standard Errors of Measurement for raw scores (dashed line) and scaled scores (solid line) on the CogAT (Form 6) Verbal Battery, Level A.

not a reasonable assumption for those who obtain very high or very low scores on the test.

Conditional errors of measurement are errors that depend on the location of a score on the score scale. The typical patterns of errors of measurement for raw and scaled scores on a fixed-length test are shown in Figure 1. As Figure 1 shows, the patterns are opposite for raw scores (i.e., number correct) than for scaled scores and other normative scores based on scaled scores (e.g., IQ scores). Differences in the patterns of errors for raw and scaled scores are caused by the way the scaling process expands the score scale at the extremes of the distribution. This means that passing or failing a single item will have a much larger effect on scale scores for those who score near the ceiling or floor of a test than for those who score near the mean. This most commonly occurs on tests in which all students in a grade are administered the same level of a test. The level of the test that is appropriate for the majority of students in a class will often be too easy for the most able students.

Tests that are scaled using Item Response Theory such as the Otis-Lennon School Ability Test (Otis & Lennon, 2003) and the CogAT (Lohman & Hagen, 2001) typically report conditional errors of measurement for scale scores. Conditional errors of measurement can be

dramatically reduced by administering a higher level of the test to more able students. For example, consider the student who receives a Verbal scale score of 221 on CogAT. Table 5.7 in Lohman and Hagen (2002) shows that the error of measurement at this score is 14.8 on Level A of the test but only 7.4 at Level D. Thus, administering the higher level of the test halves the expected error of measurement.

Differential Growth Rates

If errors of measurement were the only factor that contributed to regression to the mean, no additional regression should occur after the first retest. Suppose that only students who obtain high scores on the initial test are selected. On average, scores would be expected to decline when the students were retested. After this first retest, however, scores would regress to the mean true (or universe) score of the group—some individuals getting higher scores on subsequent retests, some getting lower scores, but the mean true score staying the same. Put differently, the correlation between the initial test score and every subsequent retest would be the same. All of these correlations would estimate the reliability of the test. However, longitudinal studies of ability do not show this pattern. Rather, the correlations tend to decrease as the interval between test administrations increases (Bayley, 1949; Humphreys & Davey, 1988). For example, the upper diagonal of the matrix in Table 2 shows correlations among Composite scores on the ITBS for 6,321 students who were tested every year from third grade to eighth grade (Martin, 1985). The lower diagonal shows correlations among IQ scores for the same intervals estimated from Thorndike's (1933) meta-analysis of 36 studies in which students were readministered the Stanford-Binet after intervals that ranged from less than a month to 5 years. The pattern in both matrices approximates a simplex: High correlations near the diagonal of the matrix decline as one moves away from the diagonal. Correlations are higher for the longer and therefore more reliable achievement test (median $r_{xx'} = .98$) than for the Binet test (estimated $r_{xx'} = .89$). The fact that correlations decline as the interval between tests increases means that factors other than error must affect retest scores on both ability and achievement tests.

Table 2
Correlations Between ITBS Composite Scores and Binet IQ Scores[a]

Grade	3	4	5	6	7	8
3		91	89	87	85	83
4	86		93	91	89	87
5	83	86		94	92	91
6	80	83	86		94	93
7	75	80	83	86		94
8	70	75	80	83	86	

Note. Decimals omitted.
[a]Above the diagonal, from Martin, 1985; below the diagonal, from Thorndike, 1933.

One possibility is differential rates of growth. A simplex pattern for correlations will be obtained as long as true-score gains are not perfectly correlated with the true-score base (Humphreys & Davey, 1988). Put differently, year-to-year gains do not have to be random, as some have hypothesized (Anderson, 1939). Rather, they only need to vary across individuals. There is in fact considerable evidence that students show different patterns of growth on ability tests. For example, McCall, Appelbaum, and Hogarty (1973) investigated changes in Stanford-Binet IQ scores for 80 middle-class children who were given the same test 17 times between ages 2½ and 17. IQ profiles for 67 of the 80 children could be classified into one of five groups. The largest group showed a slightly rising pattern of scores over childhood. Other groups showed patterns of sharp declines or increases at different ages. In general, major shifts occurred most frequently at ages 6 and 10. Note that changes in IQ reflect changes in rank within successive age groups rather than changes in ability to perform tasks. IQ scores decline even if ability improves, but at a slower rate than age-mates who obtained the same initial IQ score.

Students' growth on both ability and achievement tests from year to year is affected by maturation, interest, quality of instruction, out-of-school experiences, and many other personal and social factors. For example, instruction that engages and appropriately challenges a student can result in cognitive growth that is larger than expected.

However, the same student may be placed in a classroom with many distractions in the subsequent year and thus show less growth.[3]

Although growth rates vary across individuals, the stability of individual differences in scores that average across tasks and domains is substantial. Indeed, Humphreys (1985) estimated that between the ages of 9 and 17, true scores on a test of general ability would correlate approximately $r = .965$ with true scores on the same test administered 1 year later. As he then put it:

> It becomes easy to understand the belief in a fixed intelligence when one looks only at the small difference in true score stability from year to year between an estimated [correlation of .965] and the 1.00 required by [the assumption of] a fixed intelligence. (p. 200)

Humphreys (1985) also showed that the correlation (r) between true scores across years could be estimated by r^y, where y was the number of years separating the two test administrations. Thus, the estimated correlation between true IQ scores at ages 9 and 17 is given by $.965^8 = .75$. This means that about 60% of the children whose true scores fall in the top 3% of the distribution at age 9 would *not* fall above that cut at age 17. Of course, error in both tests would lower the observed correlation and thus result in substantially less stability across time. We never know true scores, only error-encumbered observed scores. One-year retest correlations typically range from $r = .8$ to $r = .9$. If a parallel form of the test is used, then the correlation is even lower.

Changes in Score Scales

Both the magnitude and the interpretation of changes in scores are influenced by the psychological and statistical properties of the score scale. Quite commonly, the content of ability and achievement tests differs across score levels. One can reduce these effects by presenting items in a common format at all points on the scale, by checking to ensure that individual differences in items conform to a unifactor model, and by using scaling procedures that attempt to make the scale properties constant throughout its range. However,

none of these controls can guarantee that the units of the scale will indeed be uniform, especially at the extremes. For example, the fact that all items are presented in a common format does not mean that items require the same cognitive processes. Matrix tests use a common format. However, difficult items on the Progressive Matrices test require the application of rules not required on simpler items (Carpenter, Just, & Shell, 1990). Nor does the fact that a unidimensional IRT scale can be fit to the data guarantee an equal-interval scale, especially when the full scale is constructed by vertically equating overlapping tests that are administered to examinees of different ages (see Kolen & Brennan, 2004).

Changes in the Norming Population

Longitudinal changes in status are easily confounded by nonrandom loss of cases over time. Although developmental psychologists recognize this as a potential confound in their own research, many who use test scores—particularly those normed on school children—often fail to take into account the fact that a substantial fraction of low-scoring students drop out of school. Nationally, only about one third of students complete high school (Barton, 2005). Dropout rates also vary across ethnic groups, states, and grades. Dropout rates have decreased between 11th and 12th grade and increased between 9th and 10th grade (Haney et al., 2004). The upshot is that rank-within-grade cohort means different things at 12th grade than at 8th grade or at 4th grade. Because less able students tend to drop out at a higher rate than more able students, a percentile rank of 90 means better performance for 12th graders than it does for 8th graders.

Summary

For educational, psychological, and statistical reasons, test scores obtained by high-scoring students will change from year to year. This change reflects errors of measurement in the tests that are common to all and errors that are particularly severe for extreme scorers, differential growth of students from year to year, changes in the content of score scales or the tests, and systematic changes in the representative-

ness of samples on which norms are derived. How large would these changes be as the result of the combination of these factors? One way to address this question is to set a criterion for giftedness and then either estimate (from correlations) or count (from scores) the number of students at later grades who would fail to meet the criterion. This has important implications for policy, such as how best to identify those students who will continue to excel, or how frequently schools should retest to determine eligibility for TAG services.

Estimating the Size of Regression Effects From Longitudinal Studies

One of the most important limitations of most longitudinal studies in the field of gifted education is that they follow only that portion of the population identified as gifted at one point in time. A better procedure, of course, would be to follow the entire cohort of students. However, longitudinal studies in which an entire cohort of students are repeatedly administered ability tests are rare, generally dated, and more often than not, quite small. For example, the classic Berkeley Growth Study (Bayley, 1949) had only 40 children. The Fels data used by McCall et al. (1973) had 80 subjects. Correlations computed on such small samples have large standard errors. For 40 cases, the 95% confidence interval for a population correlation of $\rho = .65$ is $r = .43$ to $r = .90$. Further, the cases are often not representative of the population. Even when samples are much larger, as in the Wilson (1983) study, differential dropout and variation in sample size across occasions at best complicates and at worst seriously biases the analyses.[4]

Achievement tests, on the other hand, are often administered every year to large groups of students. If the sample is large, the data can be reweighted better to represent the population distributions of achievement. This can in significant measure control for nonrandom loss of cases over time. Large samples also mean that correlations are quite stable. Correlations among achievement test scores exhibit the same simplex structure that is observed for ability tests. This should be expected, given the high correlations between ability and achievement tests. Indeed, ability tests are probably best understood as achievement

tests that sample general reasoning abilities developed in a culture, whereas achievement tests sample those abilities specifically developed through formal schooling. The belief that ability and achievement tests measure (or ought to measure) qualitatively different constructs has inhibited the interpretation and use of both types of tests since the earliest days of testing (Lohman, in press-a).

Martin (1985) reported a longitudinal analysis of ITBS scores for 6,321 students who were tested every year from third to eighth grade. Prior to computing the correlations, Martin reweighted the data to better approximate the distribution of grade 5 Composite achievement for Iowa students. We used his correlations to estimate the percent of students who fell in the top 3% of the Reading Total, Language Total, Mathematics Total, and Composite score distributions at grade 3 who were also in the top 3% on each subsequent retest. (See Table 2 for Composite score correlations.) The estimates assume a bivariate normal distribution of each pair of test scores.[5] The results are shown graphically in Figure 2.

The greatest regression—which is largely due to errors of measurement—occurs from Year 1 to Year 2. Only about 40% of the students who had composite scores in the top 3% in third grade also scored in the top 3% in fourth grade. Note that this occurs in spite of the fact that Composite ITBS scores are highly reliable (K-R 20 $r_{xx'}$ = .98) and show substantial stability across years (r = .91 for grade 3 to grade 4). As would be expected, regression effects were greater for the Reading, Language, and Mathematics subtest scores than for the Composite score that combines them. For each of these content scores, the fallout was approximately 50% in the first year. As the figure shows, however, regression continues at a slower rate across grades. This means that regression effects reflect more than errors of measurement. By eighth grade, the correlations indicate that only 35–40% of those who scored in the top 3% at grade 3 would still score in that range.

The procedures that many schools use to identify exceptional students were not designed to cope with regression effects of this magnitude. Indeed, some use procedures that exacerbate these effects. Others, however, use procedures that, wittingly or unwittingly, reduce regression effects.

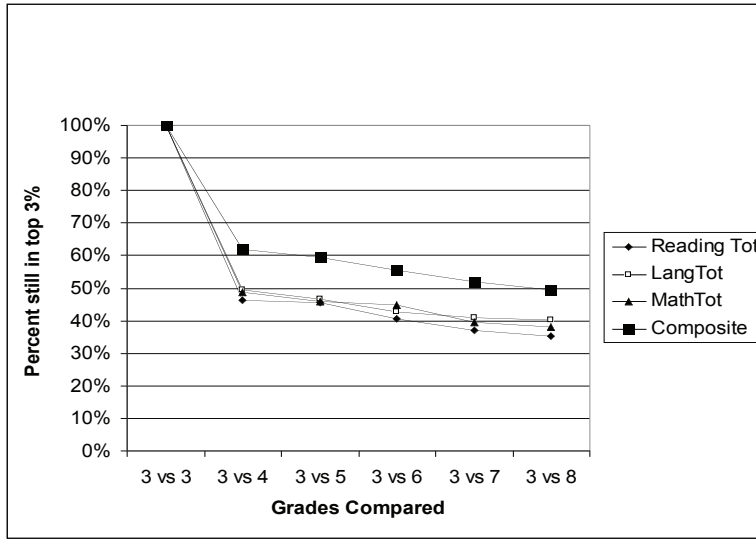

Figure 2. Percent of cases in the top 3% of the grade 3 distribution also in the top 3% of the score distributions at grades 4 through 8 for ITBS Reading, Language, Mathematics, and Composite Total scores

Regression and Common Identification Procedures

Schools use nominations, rating scales, and test scores in many different ways when selecting students for participation in special classes for the gifted. In this section, we examine some of the more common rules. The first policy is to require a high score on two or more tests. We call this the "and" rule. The second possibility is to accept a high score on either of two or more tests. We call this the "or" rule. Although rarely employed, another possibility is to average scores across two or more measures. We call this the "average" rule. The three rules are illustrated in Figure 3.

The "And" Rule

Many TAG programs set up a series of hurdles and admit only those students who surmount all of them. For example, the potential pool of applicants is first restricted to those students who are nominated by a teacher or who score above a certain score on a screening test.

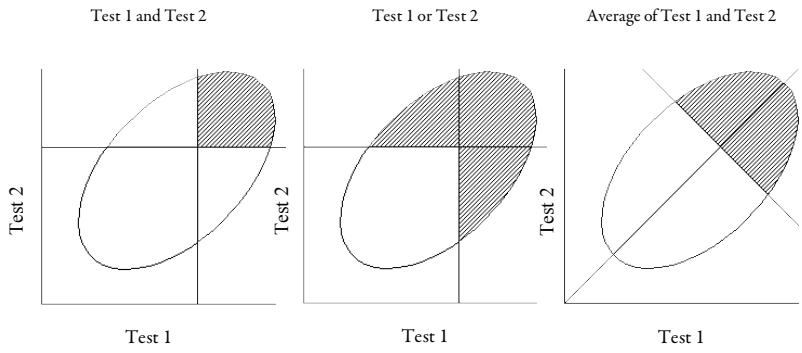

Figure 3. Plots of the conjunctive "and" rule (left panel), the disjunctive "or" rule (center panel), and the statistically optimal "average" rule (right panel).

These students are then administered a second test. Only those who exceed some score on the second test are admitted to the program.

There are both advantages and disadvantages to this procedure. The primary advantage is that it reduces the number of students who must be administered the second test. This can be important when the second test must be individually administered by a trained examiner. The second advantage of the multiple hurdles procedure is that it decreases the amount of regression that will be observed on future occasions when compared to a selection rule that uses only one test or takes the highest score on any of several tests. However, as will be shown, this effect is only observed if both tests are used to validate student classifications on the second occasion.

The primary disadvantage of the "and" rule is that the procedure assumes that the two tests are exchangeable measures of the same construct. If the tests are not exchangeable, then the sample will be biased unless a very liberal cut score is used for the first test. For example, suppose that the first "test" is a teacher nomination scale. Students who do not conform to the teacher's model of giftedness but who would have exceeded the cut score on the second test will not be considered for the program. This was one of the limitations of the Terman study (Terman & Oden, 1959). A second disadvantage

is that the selection rule is noncompensatory. A very high score on one test cannot compensate for a score on the second test that is just below the cut. Therefore, requiring students to score above a particular cut score on Test 1 *and* Test 2 restricts the number of students who are identified compared to a rule that admits on the basis of a high score on either test. But, by how much?

Table 1 shows the amount of regression to expect with various common cut scores for two selection tests that are correlated to different degrees. The table shows the proportion of students above a common cut score on both tests as the correlation between two tests varies from $r = .5$ to $r = .975$. For example, consider the case in which the common cut score is set at the top 3% and the correlation between the tests is $r = .80$. Table 1 shows that 45% of the students in the population who score in the top 3% on Test 1 are expected to score in the top 3% on Test 2. This means that 45% of the 3% who met the criterion on Test 1 or 1.35% of the total student population will be admitted when a score in the top 3% is required on both tests.

If a more lenient cut score is used for the initial nomination procedure and the same cut score is used for the final admissions test (top 3%), then the effects are much smaller. Table 1 cannot be used to estimate these effects, because it assumes a common cut score.[6] For example, once again assume that the correlation between Test 1 and Test 2 is $r = .80$. Suppose that we take the top 10% of the cases on Test 1. The top 10% on the first test includes 79% of the cases in the top 3% on the second test. An even more lenient criterion of the top 20% on Test 1 gets 93% of those who score in the top 3% on Test 2.

The policy implications are clear. If the goal is to reduce the number of students who must be administered the second test but to exclude as few of those who would obtain high scores on the second test, then one must use a lenient criterion on the screening test. This is increasingly important as the correlation between the two tests declines. If, however, both tests are equally reliable and are assumed to measure the same construct, then similar criteria can be used on both. Nevertheless, the proportion of students who clear both hurdles will be considerably smaller than the proportion who

clear either hurdle. The lower the correlation between the tests, the smaller this proportion will be. Finally, if the two tests are in fact exchangeable, then a compensatory model such as the "average" rule is more defensible.

The "Or" Rule

The disjunctive "or" rule has quite different effects. Table 1 allows one to estimate the effects of this rule, as well. As before, assume a correlation of $r = .80$ and a common cut score of the top 3%. Test 1 admits 3% of the population. Test 2 also admits 3%, but 45% of these students (as Table 1 shows) were already admitted by Test 1. The remaining 55% will be new. Therefore, 3% + (.55)(3%) = 4.65% of the student population would be admitted. Changing the rule from "and" to "or" more than triples the number of students admitted from 1.35% to 4.65%.

The disjunctive "or" rule is most defensible if the two tests measure different constructs such as language arts or mathematics. If programs (or acceleration options) are available in both domains, then one should seek to identify students who excel in either domain, not just those who excel in both domains. However, as is shown later, multiple measures of aptitude for each domain are preferred to a single measure.

The "or" is not defensible, however, when both tests are assumed to measure the same construct. For example, the test scores may represent multiple administrations of the same ability test or consecutive administrations of several different ability tests. *Error of measurement* is defined as the difference between a particular test score for an individual and the hypothetical mean test score for that individual that would be obtained if many parallel forms of the test could be administered. The highest score in a set of presumably parallel scores is actually the most error-encumbered score in that set. Therefore, unless one has a good reason for discounting a particular score as invalid, taking the highest of two or more presumably parallel test scores will lead to even more regression to the mean than would be observed by using just one score.

The "Average" Rule

If both tests measure the same construct, however, the statistically optimal rule is neither "or" nor "and" but rather "average." The "average" rule will admit more students than the restrictive "and" rule but fewer students than the liberal "or" rule. It allows for more compensation than the "and" rule but less compensation than the "or" rule. The student who has a high score on one test but a score that is just below the cut on the other test will be admitted. Essentially, students are ranked on the basis of where they fall on the 45° diagonal in the plot of scores on Test 1 versus scores on Test 2 rather than either the *X*-axis or the *Y*-axis. However, because the average of two scores will immediately regress to the mean, fewer students will meet an arbitrary cut score than will meet it if just one test is administered. With a correlation of $r = .8$, for example, 2.4% of the students would be expected to have an average score that exceeded the cut score that admitted 3% on either test alone.[6]

Regression Effects on Subsequent Retest

One of the most important considerations for any selection rule is the extent to which it effects a reasonable compromise between obtaining the most stable scores and the most valid scores. The most stable scores will generally be obtained by combining scores across different tests and occasions, with each weighted by its reliability. However, a score that averages across several domains will generally be less valid as a measure of aptitude for a specific domain than scores that capture the general and specific aptitudes needed to attain excellence in that domain. We discuss both of these issues but first focus on the stability of scores in the common scenario in which students must be nominated before they are tested. Do the admission test scores for these students exhibit greater stability than would be observed if no screening test had been administered? Intuitively, it seems reasonable to expect less regression to the mean over time, say, in IQ scores for a group of students who were first nominated by their teachers as the most able students in their class than for a group identified solely

on the basis of their IQ scores. As we shall see, however, intuitions can be wrong.

To Nominate or Not to Nominate

Simulations provide a useful method for investigating the regression effects of different decision rules when more than two variables must be considered. Here, we investigated the typical scenario in which only those students who are nominated by a teacher take an intelligence test. As already demonstrated, the number of students admitted depends on the cut scores established for the nomination procedure and the correlation between scores on the nomination rating scale and the admissions test. To simplify matters, we assume that 10% of students with the highest scores on the nomination scale are administered the intelligence test. The cut score for the intelligence test is set so that in an unselected population, 3% of the students would be admitted. For an intelligence test with *SD* of 15, this would be an IQ > 128.

Nomination procedures vary in the extent to which they measure the same characteristics as the intelligence test. In this simulation, we started with a population of 10,000 students. We varied the correlation between the nomination scale and the intelligence test from $r = .1$ to $r = .9$. A high correlation such as $r = .9$ simulates the case in which the nomination procedure is highly effective in identifying those who will obtain the highest scores on the intelligence test. The critical question is whether the nomination process reduces the amount of regression that will be seen a year later when the intelligence test is readministered. We assumed that the correlation between these two administrations of the intelligence test was $r = .8$. Table 3 shows the results.

The first column of the table shows the correlation between the nomination rating scale and the intelligence test. The second column of the table shows the number of students in a population of 10,000 students who scored in the top 10% on the nomination scale and then obtained an IQ > 128 on the intelligence test. These are the students who would be admitted to the program. When the correlation between the nomination scale and the intelligence test was $r = 1.0$,

Table 3
Effects of Nomination on Subsequent Regression to the Mean of IQ Scores

Correlation between nomination scale and intelligence test	Number of students nominated with IQ > 128[a]	Number of admitted students with IQ score > 128 1 year later[b]	Percent of admitted students with IQ > 128 1 year later[c]
1.0[d]	300	126	42
0.9	274	122	45
0.7	202	101	50
0.5	137	65	47
0.3	84	37	44
0.1	41	20	49

[a]Number of students from a population of 10,000 scoring in the top 10% on the nomination scale and the top 3% on the admissions test. [b]Number admitted scoring in top 3% on retesting; correlation between the two administrations of the admissions test was $r = .8$. [c](Column 3/ Column 2) × 100. [d]A correlation of $r = 1.0$ simulates the case in which the nomination step is omitted and all students are administered the intelligence test.

then 300 students (i.e., 3% of 10,000) would be admitted. This simulates the case in which the nomination procedure was not used and all students took the intelligence test. As the correlation between the nomination scale and the admissions test declines, many students who would have obtained IQs greater than 128 on the admissions test were excluded because they were not nominated. When the correlation is high, however, one might argue that many of the excluded students did not belong in the group in the first place and would be the students least likely to score above an IQ of 128 when the intelligence test was readministered 1 year later. The third and fourth columns of the table show that this is not the case. Although the nomination procedure reduced the number of students who were admitted, it did not significantly reduce the regression effects observed when the intelligence test was readministered 1 year later.

Note the important difference between this procedure and the case in which scores on the screening test (or nomination rating scale) and the admissions test are first combined and students are selected on the basis of their scores on the resulting composite. One of the easiest ways to combine scores is simply to sum them or average them, after first putting all scores on the same scale.[7]

Combining Ability and Achievement Test Scores

The identification of academically talented students ultimately resolves to the estimation of aptitude for rapid or advanced learning in the particular educational programs that can be offered. Aptitude is a multidimensional concept. It has cognitive, affective, and conative dimensions. The primary cognitive aptitudes for academic learning are current knowledge and skills in a domain and the ability to reason in the symbol systems used to communicate new knowledge in that domain. The primary affective aptitude is interest in the domain. The primary conative aptitude is the ability to persist in one's pursuit of excellence. Different instructional programs require or afford the use of different aptitudes. One of the most important goals for research should be to better understand the relationships between those aptitude characteristics that can be measured prior to identification and that contribute to the prediction of success in different kinds of programs.

There is much research, however, on the critical importance of the two primary aspects of cognitive aptitude for learning—prior achievement and reasoning abilities. The best way to do this is to combine scores so that they best predict subsequent achievement. When done well, both immediate and long-term regression effects will be minimized. In this section we explore some basic options for achieving this goal. To do this, we need a longitudinal data set that has both achievement and ability scores for a large sample of students.

Gustafson (2002) collected ability and achievement test scores for 2,362 students in a large Midwestern school district who were tested first in grade 4, then in grade 6, and again in grade 9. The ability test was CogAT Form 5 (Thorndike & Hagen, 1993)

Table 4
Correlations Across Grades for ITBS (Form K) and CogAT (Form 5) Scores (*N* = 2,363)

Test		Grades		
		4 with 6	6 with 9	4 with 9
ITBS				
	Reading	0.76	0.77	0.73
	Language	0.77	0.72	0.67
	Mathematics	0.74	0.73	0.67
	Composite	0.86	0.84	0.79
CogAT				
	Verbal	0.81	0.80	0.75
	Quantitative	0.75	0.77	0.71
	Nonverbal	0.72	0.74	0.68
	Composite	0.85	0.87	0.82

Note. Data from Gustafson (2002).

and the achievement test was the ITBS Form K Survey Battery (Hoover, Hieronymus, Frisbie, & Dunbar, 1993). In order to illustrate how different selection models perform over time, we looked at high achievers in two domains: *Reading* (Reading Vocabulary plus Reading Comprehension) and *Mathematics* (Mathematics Concepts, Mathematical Problem Solving, and Math Computation) at fourth grade.

Table 4 shows the correlations across the three grades for the three batteries of the ITBS and the three batteries of CogAT. In all cases, the correlation between grades 4 and 9 was smaller than the correlation between grades 4 and 6 or between grades 6 and 9.

The solid line in the left panel of Figure 4 shows the percentage of high-achieving students identified on the fourth-grade reading test who also met the same percentile-rank cut score in sixth and in ninth grade. We could not use a criterion of the top 3% because of a ceiling effect on the grade 9 tests.[8] Therefore, we selected the 7% of students with the highest scores at grade 4. As in the analyses of the Martin (1985) data, Figure 4 shows a dramatic decline in the percent of students identified at grade 4 who continued to score at or above the 93rd percentile between the first test (grade 4) and the second test (grade 6), and then a smaller decline between grades 6 and 9. The right panel of Figure 4 shows similar effects for Mathematics.

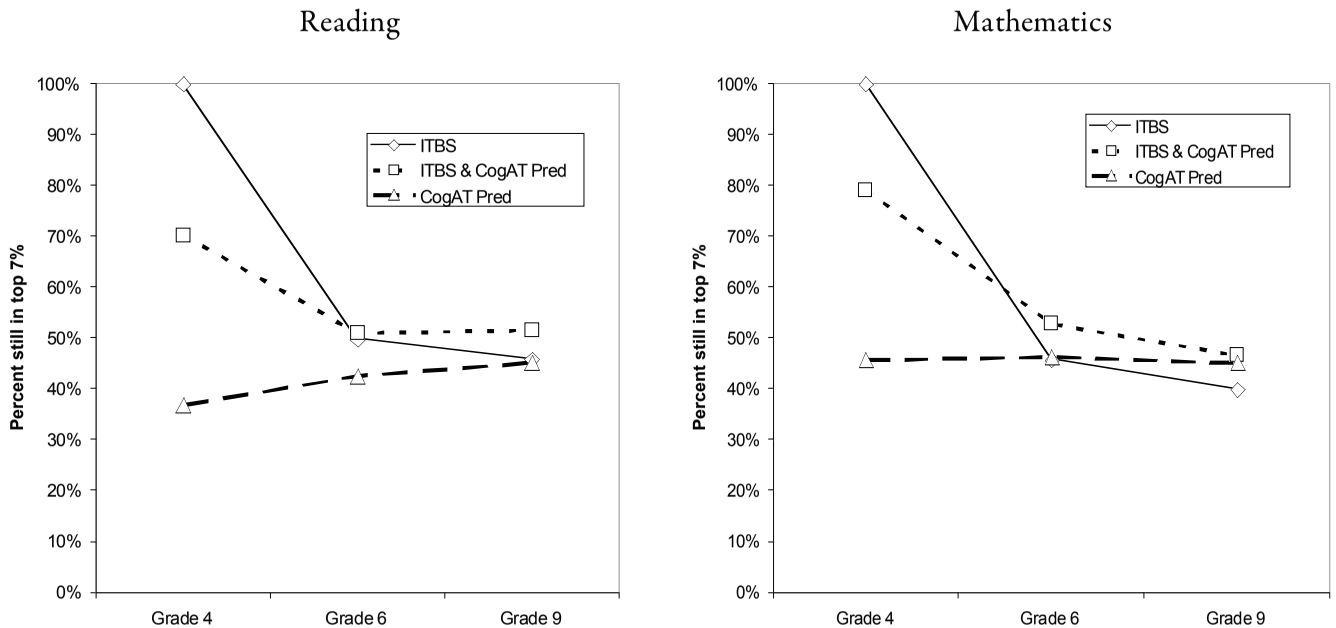

Figure 4. Students in the top 7% of the ITBS distribution who were still in the top 7% at grades 6 and 9 using three identification models. Left panel for Reading Total scores; right panel for Mathematics Total scores.

Although current achievement is a critical aspect of academic talent, it is also important to consider other characteristics that indicate readiness to continue to achieve at a high level such as reasoning abilities in the major symbol systems used in that domain, interest in the domain, and persistence. We did not have measures of interest or persistence, but did have CogAT reasoning scores in verbal, quantitative, and figural domains. Therefore, we also looked at the linear combination of the three CogAT reasoning scores that best predicted ninth-grade reading.[9] The percentage of top readers that would be identified using this estimate from CogAT scores is shown by the dashed line in Figure 4. Clearly, using predicted rather than observed reading achievement at grade 4 missed many of the best readers at grade 4. However, the figure shows that most of those who were missed did not fall in the top 7% of the reading distribution at grade 6. And by grade 9, grade 4 Reading and grade 4 CogAT scores both identified the same proportion of students who were still in the top 7% of the reading distribution. For mathematics (right panel of Figure 4), grade 4 Mathematics and grade 4 CogAT identified the same proportion of students still in the top 7% at grade 6. By grade 9, the regression estimate based on grade 4 CogAT scores identified more of those who were in the top 7% of the Math distribution than did grade 4 ITBS mathematics scores.

Because both prior achievement and reasoning abilities function as aptitudes for learning, a more effective selection model would combine current achievement and reasoning abilities in the symbol systems used to communicate new knowledge in the domain. Estimating achievement at grade 4 is straightforward. We used the child's ITBS Reading Total and Mathematics Total scaled scores. But, which of the three reasoning scores from CogAT should we use? In previous analyses of this data, we estimated the optimal weights to apply to ITBS and CogAT scaled scores at grade 4 to predict achievement in reading and mathematics at grade 9 (see Tables 2 and 3 in Lohman, 2005). These analyses showed that grade 9 Reading was best predicted by the grade 4 CogAT Verbal score, with minor contributions from the CogAT Quantitative and Nonverbal scores. Similarly, grade 9 Mathematics was best predicted by the grade 4 Quantitative score, although both the Nonverbal and Verbal batter-

ies contributed significantly, as well. Although we used the optimal weights, one can do about as well by using only the CogAT Verbal score to predict reading and the sum of all three CogAT scores (i.e., the Composite) to predict mathematics. This gave us two aptitude scores for each child in each domain: current achievement in that domain and a composite CogAT reasoning score for that domain.

Next, we combined observed achievement in fourth grade with our estimate of predicted achievement (in reading or in math) in ninth grade. Observed fourth-grade achievement and predicted ninth-grade achievement in reading or math were converted to standard or *z* scores and then summed. This ensured that both scores contributed equally to the composite. We weighted each equally because our previous analyses showed that prior achievement and reasoning abilities made approximately equal contributions to the prediction of achievement at grade 9. For detailed instructions on how to create and combine standard scores in a Microsoft Excel worksheet, see Lohman (in press-b).

The dotted lines in Figure 4 show how this selection variable performed. The largest effect was at grade 4. Although the composite score did not identify all the high scorers at grade 4, it did identify about 70% in reading and about 80% in math. At grade 6, the composite performed as well as grade 4 reading achievement alone and significantly better for math than grade 4 math achievement. By grade 9, the composite achievement-ability measure was the best predictor for reading and was about as good as CogAT scores alone for math.

As Figure 4 shows, there is a tradeoff between measurement of current achievement and aptitude for future achievement. Measures of domain-specific achievement best identify high performers at a particular point in time. However, many of these students do not continue to perform at the same stellar levels of achievement even after 1 year (see also Figure 2). On the other hand, although reasoning abilities do not identify all of the high achievers at a grade, those that they do identify are those who are most likely to continue as high achievers in subsequent years. Indeed, in mathematics at least, by grade 9 those with the highest predicted achievement based on grade 4 CogAT scores were even *more* likely to still be identified

as high achievers than those who were identified on grade 4 ITBS Math alone. The final set of analyses using both achievement and ability test scores suggests that a sensible policy for identifying talented and gifted students would combine both current achievement in particular domains and that combination of reasoning abilities that best predicts later achievement in those domains.

Policy Implications

The stability of test scores has important implications for educational policy. First, multiple scores should always be used to make educational decisions about gifted students. There are two ways that this could be done. Each student's previous test scores could be taken into account when making educational decisions, such as considering achievement test scores over the course of a few years. For example, one could look at the average of scaled scores on the two most recent assessments.[10] However, when scores are averaged, the cut score must be lowered: The more reliable average score will show some regression to the mean.

Multiple scores can also be used by combining both achievement and ability test scores that are administered at roughly the same time. Figure 4 shows that the average of ITBS achievement and the combination of CogAT scores that best predicted later achievement performed better than either measure alone in identifying those students who continued to exhibit academic excellence in particular domains. Schools should also investigate the use of measures of interest and persistence, although these measures should surely be given much less weight than measures of achievement and ability.[11] Combining scores that estimate different aptitudes needed for the development of future competence is the best way to identify talented students. However, judgments about aptitude are best made by comparing a student's scores on the relevant aptitude variables to those of other students who have had similar opportunities to develop the knowledge, skills, interests, or other attributes sampled by the assessment (see Lohman, in press-b).

Another important policy issue is the amount of time that should be allowed before students are retested for continued participation in gifted programs. Applying the entries in Table 1 to the correlations reported in Table 2 suggests that 3 years is an outside limit (2 would be better), especially if the first test is administered during the early primary grades (K–2). Finally, because test scores are especially unstable for those students with extreme scores, students who would qualify as gifted based on one test will not necessarily qualify as gifted when retested even 1 year later. Therefore, instead of using terms that imply fixed categories, such as *gifted*, perhaps educators should use words that focus less on a fixed state and instead on current accomplishment, such as *superior achievement* or *high accomplishment*.

Conclusions

Our first goal in this paper was to summarize some of the basic facts about regression to the mean for researchers and practitioners in the field of gifted education. We hoped to dispel notions that regression to the mean is attributable solely to errors of measurement. Rather, regression is determined by the correlation between two sets of scores. Anything that lowers the correlation increases regression to the mean. The data that we presented show that, even for highly reliable test scores, approximately half of the students who score in the top 3% of the score distribution in 1 year will not fall in the top 3% of the distribution in the next year. This has important implications for both research and practice.

The research implication is that we need more longitudinal investigations of individual differences in abilities of all sorts. Retesting those who are identified as gifted at one point in time provides useful information. However, as shown here, this will commonly miss many—or even most—of those who attain high scores on the attribute at some later point in time. Therefore, much more information about the origin and development of academic excellence (rather than precocity) could be obtained from studies in which the entire population of learners was followed over time.

The primary implication for practice is that one can substantially reduce the amount of regression by combining the information from multiple assessments. However, different ways of combining scores have dramatically different effects on the number of students who are admitted and the amount of regression seen in their test scores. In general, the statistically optimal method of combining similar scores is to average them.

In the end, in addition to multiple measures, local norms provide a better way to identify students for inclusion in special programs that are based in the school. Understanding that all abilities are developed and that schools play a critical role in that process can lead to policies in which children's reasoning abilities are assessed if not as regularly as their achievement, then at least at several points in their academic careers. Lacking such understanding, both selection policies and research on the gifted will continue to give mute testimony to the robustness of regression to the mean.

References

Ackerman, P. L. (1989). Abilities, elementary information processes, and other sights to see at the zoo. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology: The Minnesota symposium on learning and individual differences* (pp. 281–293). Hillsdale, NJ: Erlbaum.

Anderson, J. E. (1939). The limitations of infant and preschool tests in the measurement of intelligence. *Journal of Psychology, 3,* 351–379.

Barton, P. E. (2005). *One-third of a nation: Rising dropout rates and declining opportunities.* Princeton, NJ: Educational Testing Service, Policy Evaluation and Research Center.

Bayley, N. (1949). Consistency and variability in the growth of intelligence from birth to eighteen years. *Journal of Genetic Psychology, 25,* 165–196.

Callahan, C. M. (1992). Determining the effectiveness of educational services: Assessment issues. In *Challenges in gifted education: Developing potential and investing in knowledge for the 21st*

century (pp. 109–114). Columbus, OH: Ohio Department of Education. (Eric Document Reproduction Service No. ED344416)

Carpenter, P., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404–431.

Daniel, M. H. (2000). Interpretation of intelligence test scores. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 477–491). Cambridge, UK: Cambridge University Press.

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

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

Gagné, F., & St Père, F. (2001). When IQ is controlled, does motivation still predict achievement? *Intelligence, 30*, 71–100.

Gustafson, J. P. (2002). *Empirical Bayes estimators as an indicator of educational effectiveness: The Bryk and Raudenbush school performance model*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Haney, W., Maduas, G., Abrams, L., Wheelock, A., Miao, J., & Gruia, I. (2004). *The education pipeline in the United States 1970–2000.* Boston: Boston College, National Board on Educational Testing and Public Policy.

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

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1993). *Iowa Tests of Basic Skills, Form K: Survey Battery*. Chicago: Riverside.

Humphreys, L. G. (1985). General intelligence: An integration of actor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurement, and applications* (pp. 201–224). New York: Wiley.

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

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

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. (in press-a). Beliefs about differences between ability and accomplishment: From folk theories to cognitive science. *Roeper Review*.

Lohman, D. F. (in press-b). *Identifying academically talented minority students* (Research Monograph). Storrs, CT: The National Research Center on the Gifted and Talented.

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

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

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.

Marsh, H. W., & Hau, K. -T. (2002). Multilevel modeling of longitudinal growth and change: Substantive effects or regression toward the mean artifacts? *Multivariate Behavioral Research, 37,* 245–282.

Martin, D. J. (1985). *The measurement of growth in educational achievement.* Unpublished doctoral dissertation, University of Iowa, Iowa City.

McCall, R. B., Appelbaum, M., & Hogarty, P. S. (1973). Developmental changes in mental performance. *Monographs of the Society for Child Development, 38*(3, Serial No. 150).

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III technical manual.* Itasca, IL: Riverside.

Milich, R., Roberts, M. A., Loney, J., & Caputo, J. (1980). Differentiating practice effects and statistical regression on the Conners Hyperactivity Index. *Journal of Abnormal Child Psychology, 8,* 549–552.

Mills, J. R., & Jackson, N. E. (1990). Predictive significance of early giftedness: The case of precocious reading. *Journal of Educational Psychology, 82*, 410–419.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88,* 622–637.

Otis, A.S., & Lennon, R. T. (2003). *Otis-Lennon School Ability Test, Eighth Edition.* San Antonio, TX: Harcourt.

Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology, 94*, 3–13.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmarks Pædagogiske Institut.

Roid, G. (2003). *Stanford Binet Intelligence Scale, Fifth Edition: Technical manual.* Itasca, IL: Riverside.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Spangler, R. S., & Sabatino, D. A. (1995). Temporal stability of gifted children's intelligence. *Roeper Review, 17,* 207–210.

Tannenbaum, A. J. (1965). Review of the Culture Fair Intelligence Test. *The Sixth Mental Measurements Yearbook, 445,* 721–722.

Terman, L. M., & Oden, M. H. (1959). *The gifted at mid-life, thirty-five years follow-up of the superior child: Vol. 3. Genetic studies of genius.* Stanford, CA: Stanford University Press.

Thorndike, R. L. (1933). The effect of the interval between test and retest on the constancy of the IQ. *Journal of Educational Psychology, 24,* 543–549.

Thorndike, R. L. (1975). Mr. Binet's test 70 years later. *Educational Researcher, 4,* 3–7.

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

Tibbetts, K. A. (2004). *When the test fails: The invalidity of assumptions of normative stability in above-average populations*. Unpublished doctoral dissertation, University of Hawai'i, Mānoa.

Wainer, H. (1999). Is the Akebono school failing its best students? A Hawaiian adventure in regression. *Educational Measurement: Issues and Practice, 18,* 26–31, 35.

Willerman, L., & Fiedler, M. F. (1977). Intellectually precocious preschool children: Early development and later intellectual accomplishments. *Journal of Genetic Psychology, 131*, 13–20.

Wilson, R. S. (1983). The Louisville twin study: Development synchronies in behavior. *Child Development, 54,* 198–216.

End Notes

[1] The more general equation for predicting regression effects when the assumption of equal variance is inappropriate can be expressed in several ways. A useful equation is $\hat{Y}_p = \overline{Y} + b_{yx}\left(X_p - \overline{X}\right)$, where $\hat{Y}_p$ is the predicted Y score for person p, $\overline{Y}$ is the mean Y score, b_{yx} is the unstandardized coefficient for the regression of Y on X, X_p is the X score for person p, and $\overline{X}$ is the mean X score.

[2] Noteworthy exceptions are the dissertation by Tibbetts (2004) and Wainer's (1999) discussion of the same data. Also see the paper by Willerman and Fiedler (1977) for an example of regression in IQ scores for gifted 4-year-olds.

[3] In a recent study, Spangler and Sabatino (1995) did not observe changes in mean retest IQs for 66 gifted children in a southern Appalachian school district. However, children were excluded from the study if they "experienced remarkable sensory, physical, health-related, social, personal or family problems" (p. 208). Further, the initial test scores may have been depressed by poor educational opportunities for some of the children. The fact that the *SD* of WISC-R IQ scores more than doubled on retest supports this conjecture.

[4] Recent improvements in statistical methods for making inferences from sparsely populated data matrices offer another avenue (see, e.g., Schafer & Graham, 2002).

[5] Estimates were derived using a program called StaTable, which is available as a free download at http://www.cytel.com/statable. For bivariate normal distributions, StaTable asks for the z scores that restrict the distribution (1.8808 for the top 3%), as well as the correlation between the two measures. The proportion of scores falling in the restricted range is then given by StaTable. To determine the percentage of students falling in the top 3% upon the second measure, the proportion given by StaTable was divided by .03, the proportion falling in the top 3% at the first measure.

[6] Tables illustrating the effects of averaging test scores and of using different cut scores for one test than another had to be deleted from the manuscript to save space. These are available from the authors on request.

[7] Averaging or summing standard scores effectively weights each the same. Regression procedures allow estimation of more nearly optimal weights. However, the unit weights implied by summing scores generally function about as well as optimal weights on cross-validation as long as each score makes a reasonable contribution to the prediction.

[8] Missing one more item thus resulted in a substantial shift in percentile rank (PR). We moved down the distribution until this was no longer a problem.

[9] We used grade 9 rather than grade 4 or 6 because we were interested in predicting success over the long haul. However, using the regression weights that best predicted grade 4 or grade 6 reading would not make much difference.

[10] Note that an average of two assessments is recommended rather than a policy of requiring that the student meet the cut score on two successive assessments. The latter rule—whether applied to the same assessment administered in different years or to different assessments (e.g., achievement and ability) administered in a given year—misses many capable students.

[11] For a summary of research on the contribution of measures of motivation to the prediction of academic success, see the excellent literature review in Gagné and St Père (2001). However, the Gagné and St Père study itself probably underestimates the contribution of motivation because the major motivation variable was difference score.