ACT Research Report Series 97-2

Grade Equivalent and IRT Representations of Growth

E. Matthew Schulz
W. Alan Nicewander

ACT

May 1997

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

Grade Equivalent and IRT Representations of Growth

E. Matthew Schulz and W. Alan Nicewander

Abstract

It has long been a part of psychometric lore that the variance of children's scores on cognitive tests increases with age. This "increasing-variance phenomenon" was first observed on Binet's intelligence measures in the early 1900's. An important detail in this matter is the fact that developmental scales based on age or grade have served as the medium for demonstrating the increasing-variance phenomenon. Recently, developmental scales based on item response theory (IRT) have shown constant or decreasing variance of measures of achievement with increasing age. This discrepancy is of practical and theoretical importance. Conclusions about the effects of variables on growth in achievement will depend on the metric chosen. In this study, growth in the mean of a latent educational achievement variable is assumed to be a negatively-accelerated function of grade; within-grade variance is assumed to be constant across grade, and observed test scores are assumed to follow an IRT model. Under these assumptions, the variance of grade equivalent scores increases markedly. Perspective on this phenomenon is gained by examining longitudinal trends in centimeter and age equivalent measures of height.

Grade equivalent and IRT Representations of Growth

The use of item response theory as a model for cognitive test data has recently introduced some controversial discrepancies concerning trends in the variability of mental traits with age (Hoover, 1984a,b, Burket, 1984; 1988; Yen, 1988; Phillips and Clarizio, 1988a,b; Clemans, 1993). There is a strong, mutual reinforcement between the popular notion that variance of cognitive skills increases with age, and the fact that grade equivalent and Thurstonian scales have traditionally confirmed this trend. A trend of increasing variance is consistent with the common-sense notion that above-average students continue to develop at a faster rate than below average students. However, when IRT scales are constructed from the same or comparable data used to construct grade equivalent and Thurstonian scales, IRT variability remains constant or even decreases (Yen, 1986; Schulz, Shen, and Wright, 1990; Lee and Wright, 1992; Bock, 1983).

Differences in the growth rate of mean or median achievement also exist between metrics. The defining characteristic of a grade equivalent scale is that median achievement in the norm group increases at a constant rate of one unit per year. Thurstonian and IRT scalings of educational achievement data generally show increases in the mean to be negatively accelerated with grade (Yen, 1986; Schulz, et al., 1990; Lee and Wright, 1992).

Differences in growth trends have practical importance in research on educational achievement. In a longitudinal study of the effects of schools and other higher-level variables on change in student's educational achievement, grade equivalent and IRT

metrics led to strikingly different representations of individual differences in growth trends among students (Seltzer, Frank and Bryk, 1994). These investigators concluded that choice of metric can influence decisions about the efficacy of educational programs.

The problem of choosing a scale for research on growth in educational achievement is complicated by the arbitrary nature of scales. Educational and cognitive tests do no more than order levels of cognitive performance. One cannot pose questions about trends in variability and rates of growth until test results are put on a metric scale. The only nonarbitrary criterion of a scale is that it preserve the ordering of performance in the test data. Two scales that are equally acceptable from this perspective can lead to opposite conclusions about trends in variability and rates of growth (Braun, 1988). Zwick (1992) gives an example in which a difference of increasing variance is converted to one of decreasing variance by an order-preserving transformation.

One aim of this paper is to show how differences in IRT and grade equivalent growth trends stem from differences in the scaling models. Yen (1986) and Schulz (1990) have pointed out that grade equivalent variance is bound to increase if an alternative order preserving metric shows a pattern of constant within-grade variance and negatively accelerated growth in the mean. To demonstrate this point here, growth trends of constant variability across grades and negatively accelerated growth in the mean on an IRT (theta) metric are assumed, and test data is assumed to fit a given IRT model (see Equation (1) in the following section). A grade equivalent scale is then constructed in order to illustrate the trend of increasing variability on the grade equivalent scale.

In the standard procedure for constructing grade equivalent scales (Petersen, Kolen, and Hoover, 1989), grade equivalents have a one-to-one correspondence with true scores on grade-level tests administered to students within grade and with true scores on a scaling test administered to all students. Trends in the distribution of number correct scores on such tests will be examined in this study, but will not be used to construct the grade equivalent scale. Instead, thetas will be mapped directly into grade equivalent scores because there is a one-to-one correspondence between thetas and number correct true scores in the IRT model. The purpose of examining trends in number correct scores is to demonstrate their relationship to trends in the theta metric.

Two methods will be used to map IRT ability parameters directly to grade equivalents. One method uses quadratic regression, and is suitable when the growth in mean achievement on an IRT scale exhibits a simple quadratic trend, as will be assumed in this study. A more general, but less exact method, called integer-assignment, maps theta values to the most probable grade (integer, grade equivalent value), according to the relative density of the assumed within-grade theta distributions. The later method is considered more suitable when growth in the mean is not a simple quadratic function over grade. Both methods are expected to yield comparable results.

To add perspective on the meaning of growth trends in either metric, an analogy to growth in a physical characteristic, height, is developed. Growth in physical characteristics has long served as a model for growth in mental traits (Bloom, 1966; Bock, 1989). As will be seen, centimeter measures of height, grouped by age, show trends of decreasing, as well as increasing variance with age and nonlinear rates of growth in the

mean. These trends provide a basis for interpreting similar trends in IRT measures of educational achievement. The analogy is extended further by mapping centimeter measures of height into age equivalent scores. The relationship between age equivalent and centimeter growth trends in height is comparable to the relationship between grade equivalent and IRT growth trends in educational achievement.

Methods

Assumptions

Let Θ represent a latent scale of achievement, and let the probability of a correct answer to a multiple choice achievement test item be the following logistic function of Θ:

$$P(\theta) = 0.2 + \frac{0.8}{1+\exp(1.7(b-\theta))} . \qquad (1)$$

b is the Θ-coordinate of the point of inflection of the regression line of $P(\theta)$ on θ for the given item.

Let j represent grade, and let the distribution of student achievement within grade j be $N(\mu_{j\theta},\sigma_{j\theta}^2)$, where

$$\mu_{j\theta} = -.133 + \frac{35*j - j^2}{30} \qquad j=1,2,...,12. \qquad (2)$$

and $\sigma_{j\theta}^2=1$ for all j. The within-grade mean and standard deviation of Θ are plotted by grade in Figure 1. The negatively accelerated rate of growth in the mean in Equation (2) is a reasonable approximation to observed trends (Yen, 1986; Schulz, et al., 1990; Lee, et al., 1992). Constant within-grade variance across grades is also an approximation to

reported IRT trends (Schulz, et al., 1990; Lee, et al., 1992). Marked decreases in IRT variability over grade (Yen, 1986) are not taken into account here because they may have been due, in part, to problems with estimation procedures (Williams, et al., 1995).

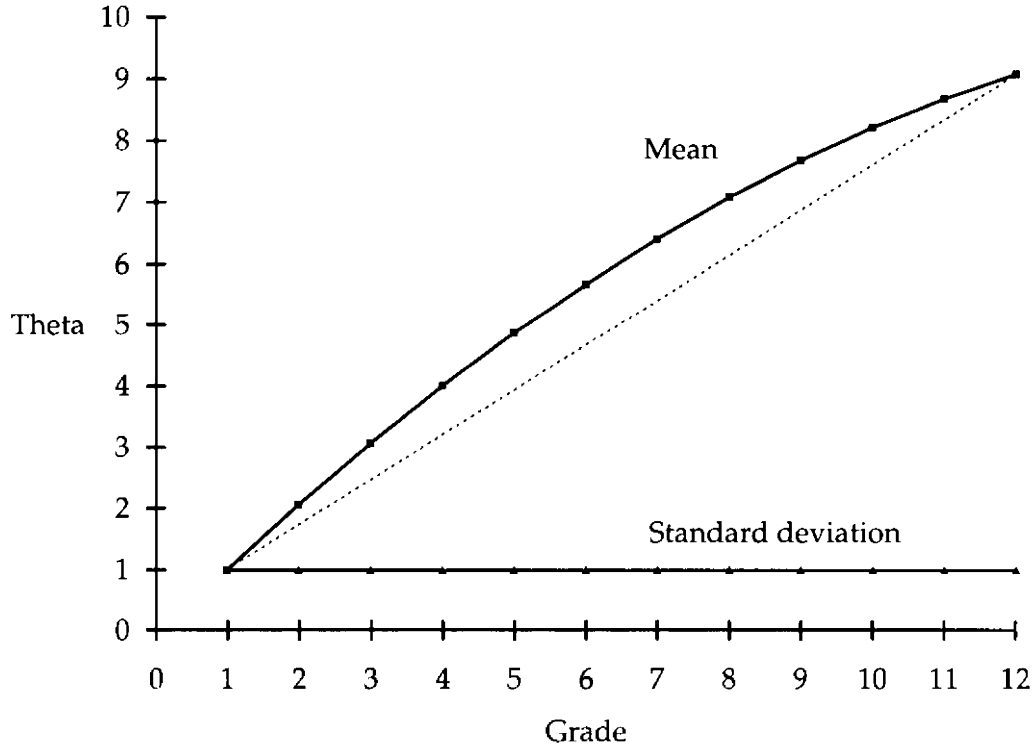

Figure 1. Hypothetical trends in the mean and standard deviation of a latent educational achievement variable (Θ). The dashed line shows linearity for purposes of comparison to the nonlinear trend in the mean.

Test Specifications and Trends in Number Correct Scores

Grade-level Tests. Let the grade-level test for grade j consist of n items with $b=\mu_{j\theta}$, let $P_j(\theta)$ be the probability that a student with $\Theta=\theta$ gets a grade-j item correct, let S represent the number correct score on the grade-level test, and let $f(\theta \mid j)$ represent the

theta density function within grade j. The within-grade mean of S is constant across grades:

$$E[S|j] = \mu_{jS} = \mu_S = n \int_{-\infty}^{\infty} P_j(\theta) f(\theta|j) d\theta \quad . \tag{3}$$

Constancy in the μ_{jS} over grades is due, in this case, to the relative difficulty of the grade-level tests being the same across grades ($b=\mu_{j\theta}$), and to the assumption that the within-grade distributions of Θ are identical across grades except for the mean. The within-grade variance of S is also constant across grades:

$$Var(S|j) = \sigma_{jS}^2 = \sigma_S = E_\Theta[Var(S|\theta)] + Var_\Theta(E[S|\theta]) \tag{4}$$

where

$$Var_\Theta(E[S|\theta]) = \int_{-\infty}^{\infty} (n P_j(\theta))^2 f(\theta|j) d\theta \quad - \quad \mu^2(S) \tag{5}$$

and

$$E_\Theta[Var(S|\theta)] = n \int_{-\infty}^{\infty} P_j(\theta)(1-P_j(\theta)) f(\theta|j) d\theta \quad . \tag{6}$$

σ_{jS}^2 is constant across grades for the same reasons μ_S is constant across grades. Given the assumption of model fit, normally distributed Θ within grade, and $b=\mu(\theta|j)$ for grade-j items, the trend in σ_{jS}^2 versus grade will have the same sign as the trend in $\sigma_{j\theta}^2$ versus grade. Both trends have zero slope in the present case due to the defined condition $\sigma_{j\theta}^2=1$ for all j.

Scaling Test. Let the scaling test consist of k items from each of the grade-level tests, and let M represent the number right score on the scaling test. The within-grade mean of M is:

$$E[M|j] = \mu_{jM} = k \int_{-\infty}^{\infty} [\sum_{j=1}^{12} P_j(\theta)] f(\theta|j) d\theta \ . \tag{7}$$

The within-grade variance of M, σ_{jM}^2, follows the general form of Equation (4), where

$$Var_{\Theta}(E[M_j|\theta]) = \int_{-\infty}^{\infty} [\sum_{j=1}^{12} (kP_j(\theta))^2] f(\theta|j) d\theta \ - \ \mu_{jM}^2 \tag{8}$$

and

$$E_{\Theta}[Var(M_j|\theta)] = k \int_{-\infty}^{\infty} [\sum_{j=1}^{12} P_j(\theta)_j (1-P_j(\theta))] f(\theta|j) d\theta \ . \tag{9}$$

Figure 2 shows a plot of μ_{jM} and σ_{jM}^2 when k=10. There is a slight S-shape in the plot of μ_{jM}, and a bell-shape to the trend of σ_{jM}^2. Both trends are connected to the fact that the difference between the number correct scores of any given pair of students is relatively small when a test is very easy or very hard for both students. Students for whom a given test is very easy or very hard are said to be performing near the test's ceiling or floor. Floor and ceiling effects of the present scaling test include the apparent shrinkage in within-grade variability at upper and lower grades, and the decline in the difference between means of adjacent grade groups at upper and lower grades. With the help of such effects, and the right set of test specifications, one can create virtually any trend in the distributions of number correct scores.

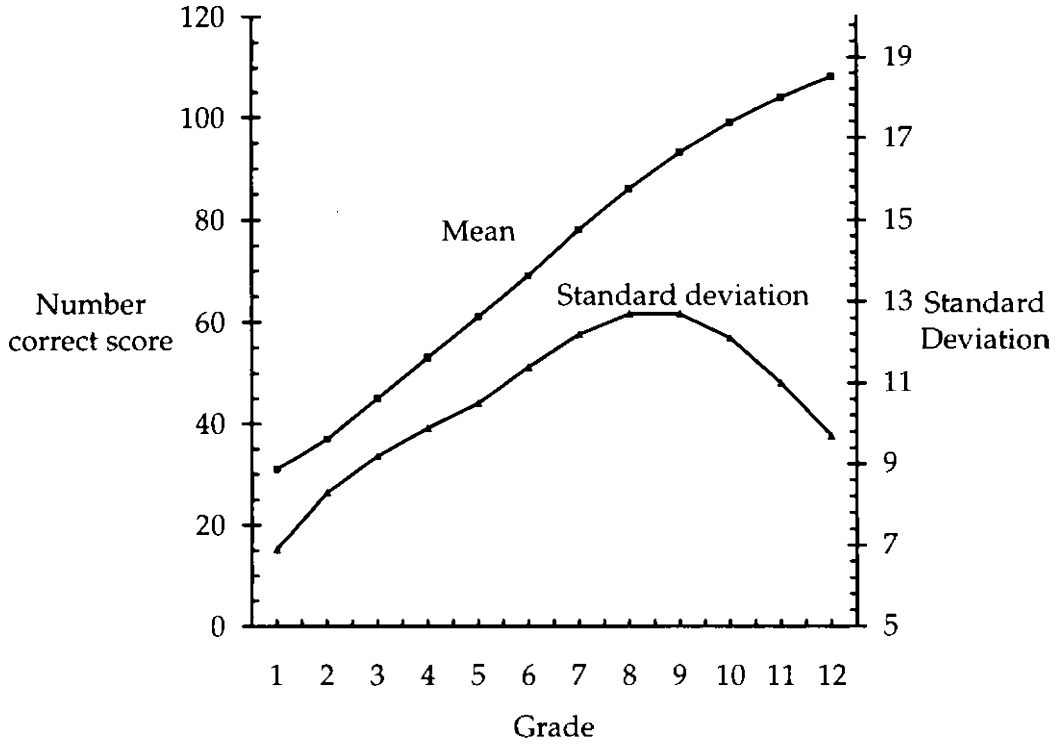

Figure 2. The mean and standard deviation, by grade, of the number correct score on a hypothetical scaling test administered to all students (see text for test specifications).

Mapping IRT Values Into Grade Equivalents

Let G denote the grade-equivalent variable, let $H_j(g)$ be the cumulative density of grade-equivalent scores within grade j, and let $F_j(\theta)$ be the cumulative density of Θ within grade j. Because true scores on the scaling test are a one-to-one function of Θ, and Θ is normally distributed within grade, $H_j(g)$ is defined for g=1,...,12, exactly as in the true score procedure (Petersen, et al., 1989) by

$$H_j(g) = F_j(\mu(\theta|j')), \qquad j'=g=1,2,...,12; \quad j=1,...,12 \quad . \tag{10}$$

That is, the grade-j percentile ranks of grade equivalent values g=1,...,12 are, respectively, the grade-j percentile ranks of the grade j' median θs, where j'=1,...,12. The points labeled "Medians" in Figure 3 are consistent with Equation (10). These "median-by-definition" anchor points define j as the median grade-equivalent value for grade j if Θ is mapped directly to G.

The use of interpolation between the median anchor points in Figure 3, to map Θ to G, is equivalent to mapping true scores on the grade-level tests to *G* as described by Petersen, et al., (1989). The procedures are equivalent because grade-level test true scores are one-to-one transforms of Θ.

Additional anchor points for mapping Θ to *G* are labeled "equal-density" in Figure 3. These were computed as follows: Let $\theta_{j+.5}$ represent the θ for which $f(\theta \mid j) = f(\theta \mid j+1)$. If the within-grade distribution of Θ is normal with variance constant across grade, then $\theta_{j+.5} = (\mu_{j\theta}+\mu_{(j+1)\theta})/2$. That is, $\theta_{j+.5}$ is exactly half-way between $\mu_{j\theta}$ and $\mu_{(j+1)\theta}$. If within-grade variances are not equal, but Θ is normally distributed within each grade, then $\theta_{j+.5} = (W_j\mu_{j\theta} + W_{j+1}\mu_{(j+1)\theta})/(W_j+W_{j+1})$, where W_j and W_{j+1} are the within-grade standard deviation of Θ for, respectively, grades *j* and *j*+1. The open squares in Figure 3 represent the equal-density points. These have the coordinates $(\theta_{j+.5}, g=j+.5)$, *j*=1,...,11. From visual inspection, these points are in the same trend line as the median-by-definition anchor points.

Rather than interpolating between the anchor points in Figure 3, two analytical methods were used to map Θ into grade equivalents.

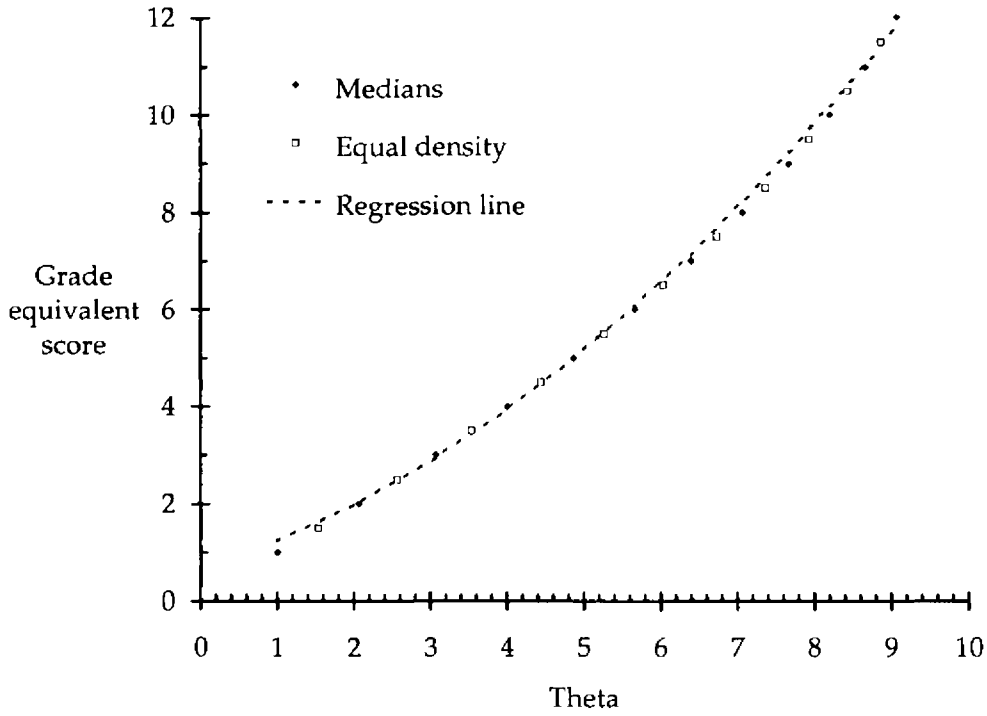

Figure 3. Relationship between grade equivalent scores and achievement on the Θ metric. Coordinates of points designated 'medians' are the median grade equivalent and median Θ within grade. Coordinates of points designated 'equal density' are the lower median grade equivalent plus 0.5, and the Θ value that is equally likely to correspond to the lower or higher grade. The quadratic regression line was estimated using both 'median' and 'equal density' points.

Mapping through Quadratic Regression. A quadratic regression of G on Θ was fit to the anchor points plotted in Figure 3. Median-by-definition anchor points, $(\mu_{j\theta}, g=j)$, and equal-density anchor points, $(\theta_{j+.5}, g=j+.5)$, yielded identical regression equations. The regression equation was quadratic with positive acceleration:

$$G = 0.57 + 0.5\theta + 0.08\theta^2 \ . \qquad \textbf{(11)}$$

This equation fit the points in Figure 3 very well (R^2=.999). Let Equation (11) be

expressed in the following general form:

$$G = \beta_0 + \beta_1\theta + \beta_2\theta^2 \ . \tag{12}$$

Then the mean of G for any grade, j, is:

$$\begin{aligned} E[G|j] = \mu_{jG} &= E[(\beta_0 + \beta_1\theta + \beta_2\theta^2)|j] \\ &= \beta_0 + \beta_1 E[\theta|j] + \beta_2 E[\theta^2|j] \\ &= \beta_0 + \beta_1\mu_{j\theta} + \beta_2(\mu_{j\theta}^2 + \sigma_{j\theta}^2) \ . \end{aligned} \tag{13}$$

The last line of (13) follows from the identity $\sigma_{j\theta}^2 = E[\theta^2|j] - \mu_{j\theta}^2$. By using a Taylor series expansion, it can be shown that the grade-j variance of G is:

$$\begin{aligned} \sigma_{jG}^2 &= E[G^2|j] - \mu_{jG}^2 \\ &= E[(\beta_0 + \beta_1\theta + \beta_2\theta^2)^2|j] - \mu_{jG}^2 \\ &\ \ \vdots \\ &= 2\beta_2^2\sigma_{j\theta}^4 + \sigma_{j\theta}^2(\beta_1 + 2\beta_2\mu_{j\theta}^2) \end{aligned} \tag{14}$$

The grade-j median of G can be expressed as:

$$\text{Med}(G|j) = \beta_2^2\mu_{j\theta} + \beta_1\mu_{j\theta} + \beta_0 \ . \tag{15}$$

Equations (11) and (14) exhibit the functional relationship between trends in grade equivalent and theta variance when the rate of increase in mean Θ is a negatively accelerated function of grade. First, Equation (11) shows that the regression of grade equivalents (G) on Θ will have a positive β_2 coefficient (.08 in this case). Second, Equation (14) shows that if β_2 is positive, grade equivalent variance (σ_{jG}^2) is bound to increase more than theta variance ($\sigma_{j\theta}^2$) (assuming there is no decrease in the mean of

Θ ($\mu_{j\theta}$)). It is conceivable that σ_{jG}^2 could decrease if $\sigma_{j\theta}^2$ were to decrease enough to offset the magnitude of β_2 and the magnitude of change in $\mu_{j\theta}$. On the other hand, σ_{jG}^2 could *increase* even if there were a *decrease* in $\sigma_{j\theta}^2$.

The within-grade mean, median, and standard deviation of grade equivalents were computed using the results of Equations (11) to (15) and the a priori values of $\mu_{j\theta}$ (Equation (2)) and $\sigma_{j\theta}^2$=1 for all j.

Mapping through Integer Assignment. In integer assignment mapping, the equal-density points on the Θ scale were boundaries for open intervals within which all thetas were mapped to the most probable grade. For example, thetas between $\theta_{3+.5}$ and $\theta_{4+.5}$ were mapped to a grade (grade equivalent value) of 4. Thetas below $\theta_{1+.5}$ were mapped to a 1 (i.e., first grade). Thetas above $\theta_{11+.5}$ were mapped to 12. These limits (1 and 12) on grade equivalent values were considered too restrictive for grades below 3 or above 8. Grade equivalent distributions were therefore estimated only for grades 3 to 8.

For computing the within-grade mean and variance of grade equivalents, the weight assigned to a given integer (grade equivalent) was based on the area of the within-grade theta distribution over the interval mapped to the given integer. The median grade equivalent for grade j was computed as a continuous value using standard methods of interpolation.

Results

Before describing the grade equivalent trends, it is important to evaluate and compare the performance of the mapping methods. Values for grade equivalent means, medians, and standard deviations by grade and mapping method are shown in Table 1. Both methods produced medians that were close to the value they would have had if the true score procedure of obtaining grade equivalents had been used.

TABLE 1

Grade equivalent score distributions by grade and method

	Method of mapping thetas to grade equivalents					
	Integer-Assignment			Quadratic Regression		
Grade	Median	Mean	Std. Dev.	Median	Mean	Std. Dev.
1	Not applicable			1.2	1.2	.68
2				2.0	2.0	.85
3	3.0	3.0	1.1	2.9	3.0	1.0
4	4.0	4.1	1.2	3.9	4.0	1.2
5	5.0	5.1	1.3	4.9	5.0	1.3
6	6.0	6.1	1.4	6.0	6.1	1.4
7	7.0	7.1	1.5	7.1	7.2	1.5
8	8.0	8.1	1.6	8.2	8.2	1.7
9	Not applicable			9.2	9.2	1.7
10				10.1	10.2	1.8
11				11.0	11.1	1.9
12				11.8	11.8	2.0

With integer assignment, the median for a given grade, *j*, was within .03 of the intended value, *j*; with quadratic regression, the difference was no larger than .2. The median trend plotted in Figure 4 is the intended, linear trend.

The mapping methods also agreed closely with each other. As shown in Table 1, the standard deviations obtained by the method of integer-assignment were within .1 of those obtained by quadratic regression--the average absolute difference was only .04. Mean values were within .1 of each other, and median values within .2.

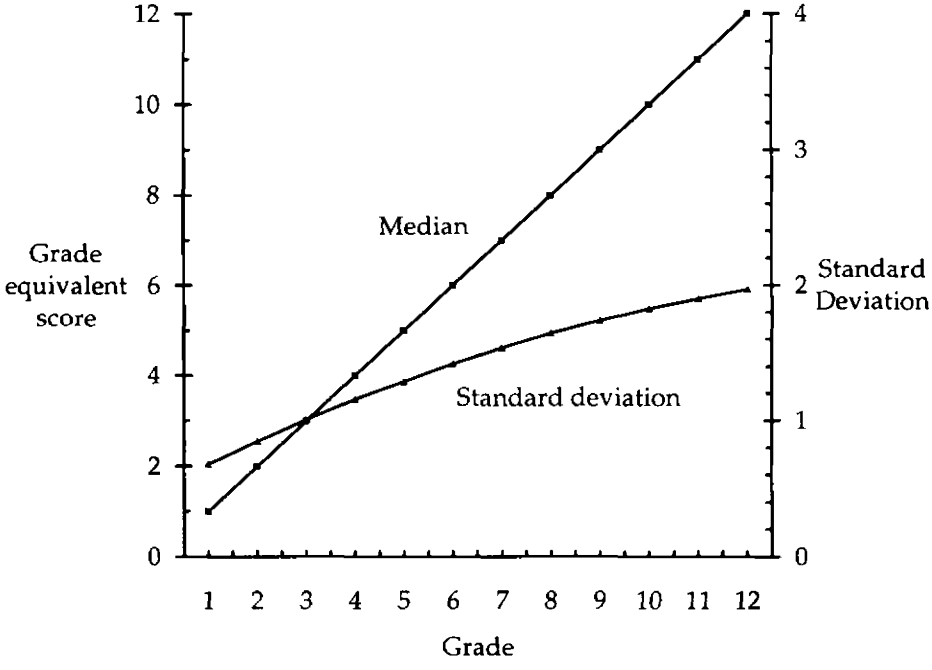

Figure 4. Trends in the median and standard deviation of grade equivalent measures.

According to both methods of mapping, the within-grade variability of grade equivalent scores increased approximately 1.6-fold from grade 3 to 8. Over grades 1 to 12, to which only the quadratic regression method of mapping was applied, the within-

grade standard deviation of grade equivalents increased approximately 3-fold, from .68 to 2.0. This trend is illustrated in Figure 4.[1]

Height Analogy

Trends in Centimeter Measures

Height data were obtained from the Fels Longitudinal Growth Study (Wright State University School of Medicine, Division of Human Biology). The data consisted of the height in centimeters of 212 boys on whom a total of 6,605 measures of height were made between the ages of 2.75 and 18.7 years. After editing, the data included the height of 160 boys measured within .1 year of their 3rd through 18th birthdays. Longitudinal trends in the mean and standard deviation are plotted in Figure 5, and corresponding values are given in Table 2.

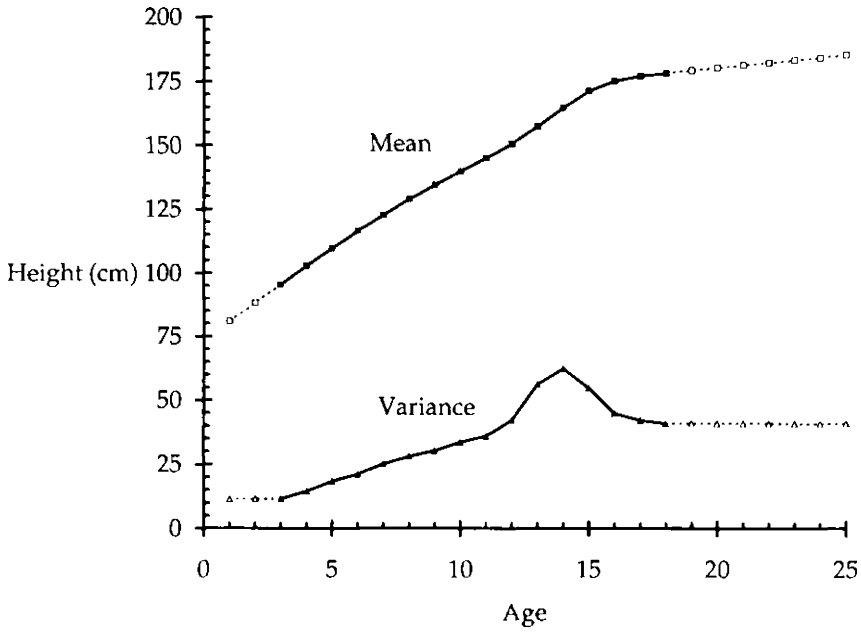

Figure 5. Trends in the mean and variance of the height in centimeters. Solid lines are based on Fels data. Dashed lines are extrapolations for purpose of constructing age equivalent measures.

Certain phenomena that may be considered applicable to cognitive growth, seem to account for complex trends in both the mean and variability of height. A collective growth spurt appears to start at age 10 and last until about age 14. This accounts for positive acceleration in mean height from ages 10 to 14. It seems likely that individual differences in the onset of this growth spurt contribute to the large increase in variance over this same period. Variance in height is maximum at age 14 because some boys have not yet begun their growth spurt, while other boys have reached full adult height. After age 14, the rate of growth in the mean is negatively accelerated with age and variance decreases as late-starters catch up with those who have reached their maximum height.

Similarly, certain cognitive skills, such as reading, could exhibit a peak in within-grade variance during early grades due to large individual differences in the onset of development. Many parents teach their children to read two or three years earlier than they would otherwise learn in school. This early advantage, however, does not necessarily persist into later primary grades, and thus, formal schooling could cause a decrease in variance of reading achievement, as measured by multiple choice test questions, over time. Other cognitive traits, such as mathematics skill, might not exhibit the same trends either because fewer parents teach their children mathematics or because the age at which achievement, as measured by multiple choice items, begins to level off might be much later for a skill like mathematics than for reading.

Trends in Age Equivalent Measures

Centimeter height was mapped to age equivalent height using the method of

integer-assignment. This method required extrapolation of centimeter height data for ages 1, 2, and 19 to 25. In order to estimate points of equal density between adjacent ages, and to assign age-specific weights to age equivalent values (for computing age equivalent means, variances, and medians by age), centimeter height was assumed to be normally distributed at each age, with means and standard deviations shown in Table 2 (extrapolated data is not shown). Heights below the equal density point for ages 1 and 2 were mapped to 1. Heights above the equal density point for ages 24 and 25 were mapped to 25.

TABLE 2

Distribution of height by age and metric

	Metric			
	Centimeters		Age equivalents	
Age	Mean	Std. Dev.	Median	Std. Dev.
3	95.4	3.4	3.0	.57
4	102.7	3.8	4.0	.61
5	109.7	4.3	5.0	.69
6	116.4	4.6	6.0	.76
7	122.7	5.0	7.0	.85
8	128.8	5.3	8.0	.95
9	134.5	5.5	9.0	1.0
10	139.8	5.8	10.0	1.1
11	145.0	6.0	11.0	1.1
12	150.5	6.5	12.0	1.1
13	157.3	7.5	13.0	1.2
14	164.7	7.9	14.0	1.8
15	171.2	7.4	15.1	3.0
16	175.1	6.7	16.1	3.6
17	177.1	6.5	17.1	3.8
18	178.1	6.5	18.0	3.9

Median age equivalent height was within .1 of the corresponding age, as shown in Table 2. As shown in Figure 6, growth in the median is practically linear with age, as expected. The standard deviation of age equivalent height increases slowly to age 13, then increases dramatically.

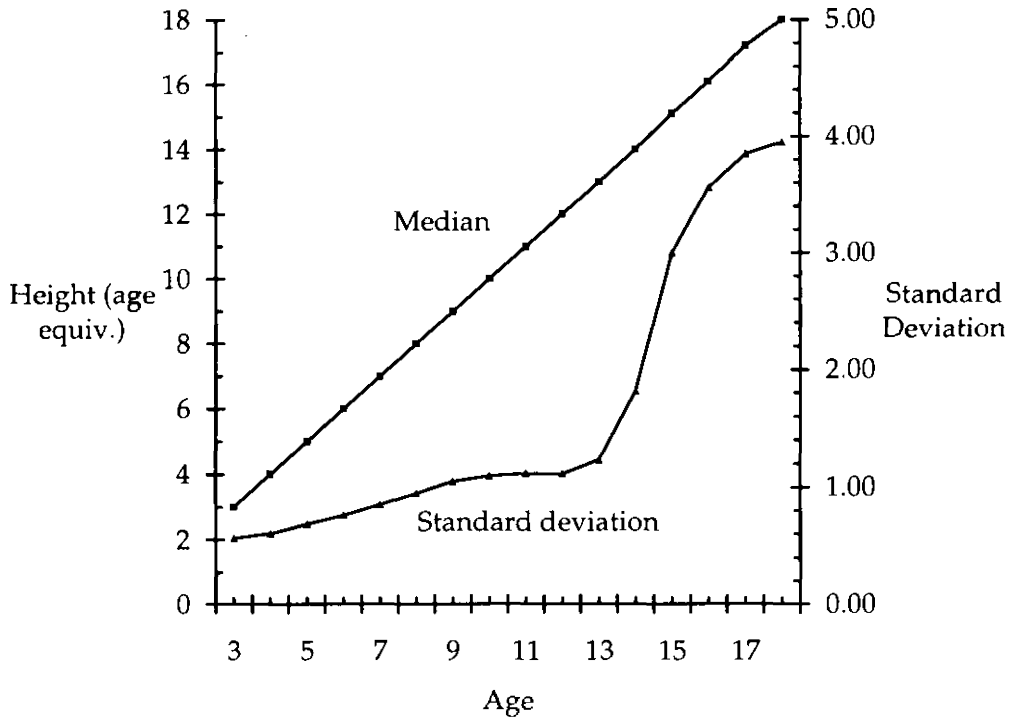

Figure 6. Trends in the median and standard deviation of age equivalent measures of height.

Discrepancies between age equivalent and centimeter trends in variability are related to nonlinear growth in the mean (in centimeters). When growth in the centimeter mean is linear (ages 3 to 10), both metrics show about the same increase in standard deviation: the age equivalent standard deviation increases nearly two fold (5.8/3.4) as does the standard deviation in centimeters (1.1/.57). When growth in the centimeter mean is positively accelerated (ages 10 to 13), the standard deviation in age

equivalents increases less than the standard deviation in centimeters (1.2/1.1 versus 7.5/5.8). When growth in the centimeter mean is negatively accelerated (ages 14 to 18), the standard deviation in age equivalents *increases* (1.8 to 3.9), even though the standard deviation in centimeters *decreases* (7.9 to 6.5).

Discussion

This paper provides a compelling demonstration of the arbitrary nature of growth trends in cognitive variables. Two metrics, both of which preserve the order of performance levels in test data, produced different pictures of cognitive growth. The differences were seen to arise strictly from differences in the scaling models. Time-indexed measures (by grade or age) will show an inflated rate of increase in variance over time relative to an alternative, order-preserving metric that shows negative acceleration in the conditional mean over time. From this demonstration, one should not expect growth trends in different metrics to look the same. Growth trends in different metrics mean different things. It falls to the investigator to carefully consider the meaning of scale units and to select the scale that gives growth trends the most useful meaning for the problem at hand.

The unit on the grade equivalent scale is defined by the indexing of performance levels on the test to grade levels. Performance levels are indexed to the grade and month of the school year at which the level of performance (on the test taken by the student) is typical. This indexing system conveys more meaning more clearly to parents and teachers of students, particularly at the elementary grades, than probably any other type of scale (Hoover, 1984a). It seems reasonable to suppose that trends in growth and

variability on the scale may also have practical use in some contexts.

On other grounds, researchers may take exception to the fact that the grade equivalent method of indexing forces the median rate of growth in the norm group to be linear. Schulz (1990) argues that a scale is not suitable for studying growth if it involves making a priori assumptions about the shape of growth. A scale must be free to detect variation in the onset, duration, and intensity of critical periods of growth, and the attainment of an asymptote, as were seen in this study with height. These phenomena, like the notion of increasing variance with age, are plausible and intuitively compelling. This is not to say that a scale should be preferred because it exhibits such features. Only that a scale must be free to exhibit such features.

Another basis on which researchers may find a problem with time-indexed scales is shown again by the analogy to height. There were two ways that the variance of age equivalent height increased: 1) when age-conditional means of centimeter height became more alike (after age 14), and 2) when the variance of centimeter height increased (from ages 3 to 10). Likewise, the variance of grade equivalent measures of educational achievement can increase by, 1) grades becoming more alike in the behaviors represented by the test, and 2) students within grades becoming more different in the behaviors represented by the test. In other words, between-group differences are confounded with within-group differences. Educational researchers interested in growth will want to be aware that time-indexed measures have this potential for confounding.

The key arbitrary scaling convention in IRT is that the correct response to any test item is a function of achievement (Θ). This assumption specifies a two-way

correspondence between numerical scale values (Θ) and the empirical observations (item responses) of the property being measured. This correspondence is one of the criteria for representational measurement. Yen (1986) discusses some distinctions between representational and index measures (grade equivalents being an index measure) that might be of interest to researchers choosing a scale for assessing cognitive growth. An illustration of the distinction between these kinds of measurement in this study, for example, is that test data and trends in grade-equivalent scores could be generated from assumptions about Θ; neither test data nor trends in the Θ metric can be derived from assumptions about grade equivalent scores.

The particular form of the item response function is also arbitrary. Lord (1980, p84) argues that the logistic function does not necessarily make the Θ metric more desirable than other functions. He gives an example of a monotonic transformation of Θ to Θ^*. The item response function on Θ^* is not logistic, but is simple and interpretable. The transformation would have no effect on data-model fit or on the representational potential of the IRT model, but growth trends in the Θ^* metric would look quite different from those in the Θ metric. Thus, the logistic function is a key ingredient for the shape of growth trends when the model is applied to real data. It does not, however, determine the shape of growth trends independently of the data. This is an important distinction from the a priori linear growth rate of time-indexed scales.

It seems reasonable to suppose that trends in the log-odds of success, like grade equivalent trends, could be useful and practical for some purposes. A given amount of change on the Θ scale means there is a corresponding change in the log-odds of success

on any given item calibrated to the scale. This study showed that there is a conditional relationship between trends in the within-grade variance of number correct scores (for on-level tests) and trends in the within-grade distribution of Θ. Increasing variability on a Θ scale means that differences are increasing among students in terms of their log-odds of success on items calibrated to the scale. This correspondence between the scale and test data could be an appropriate basis for conclusions about educational programs and achievement, particularly when test items sample a criterion domain of educationally or socially significant behaviors.

Based on the demonstration provided by this study, we recommend that when discrepancies between growth trends emerge with real data, investigators consider whether differences between models, as opposed to estimation problems and technical faults could account for the discrepancies. The discrepancies noted between variance trends in grade equivalent and IRT metrics (Schulz, et al., 1990; Lee and Wright, 1992) are exactly what one would expect if the Θ growth rate of the norm group for the tests used in these studies were negatively accelerated. [Growth in the norm group, but not necessarily the study group, would have to be negatively accelerated because the grade equivalent scores for the study group were norm-referenced.] This seems likely since Thurstonian and IRT growth rates for other standardized test batteries are negatively accelerated (Yen, 1986), and the mean growth rates for the study groups themselves were also slightly negatively accelerated (Schulz, et al., 1990; Lee and Wright, 1992).

Continued research and refinements of IRT methods and theory are needed to tease out how Θ growth trends depend on stage of development, skill, item bank, type

of IRT model, and estimation procedure. Trends of *decreasing* variability in Θ may be partially a property of estimation methods (Williams, et al., 1995; Omar, 1996), as opposed to a property of the IRT model used. When the within-grade population variance of Θ on the NAEP mathematics subtest was estimated directly rather than relying on estimated thetas (Camilli, Yamamoto, and Wang, 1993) it increased from grade 4 to 8, but decreased from grade 8 to 12. Differences between IRT models may also be a factor. The within-grade variability of one-parameter IRT measures of reading decreased only slightly across primary grades (Schulz, et al., 1990; Lee and Wright, 1992), and the variability of similar measures of mathematics achievement remained constant (Lee and Wright, 1992). Becker and Forsyth (1992) found that the within-grade variability of one-parameter IRT, three-parameter IRT, and Thurstonian measures of performance on an ITED vocabulary test all increased across grades 9 to 12.

In summary, growth trends based on cognitive test scores are fundamentally arbitrary because these scores are ordinal. Since ordinal measurement scales allow such a large variety of transformations, compared to metric measurements, one should expect to find different shapes of growth functions across time, depending on the scale used--and, as we have tried to demonstrate in this inquiry, one should also expect different patterns of variances across time depending on the measurement scale. Some cognitive scales will show the increasing-variance-with-age trend; other measurement scales for the same trait (or even the same test) can be expected to indicate a decreasing-variance-with-age trend--or even a constant-variance-with-age pattern. As long as mental traits are measured with scores that only rank-order persons, it may well be impossible to

determine the 'true' relationship between the age of children and the amount of variability in their cognitive performance.

References

Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement*, **29**(4), 341-354.

Bloom, B. S. (1966). *Stability and Change in Human Characteristics*. New York: John Wiley & Sons, Inc.

Bock, R. D. (1983). The mental growth curve reexamined. In D. J. Weiss (Ed.), *New horizons in testing* (pp 205-219). New York: Academic Press.

Bock, R. D. (1989). Prediction of Growth. In L. M. Collins & J. L. Horn (Eds.), *Best Methods for the Analysis of Change* (pp 126-136). Washington, DC: American Psychological Association.

Braun, H.I. (1988). A new approach to avoiding problems of scale in interpreting trends in mental measurement data. *Journal of Educational Measurement*, **25**(3), 171-191.

Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues and Practice*, **3**, 15-16.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, **17**(4), 379-388.

Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment*, **1**(4), 329-347.

Hoover, H. D. (1984a). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice*, **3**, 8-14.

Hoover, H. D. (1984b). Rejoinder to Burket. *Educational Measurement: Issues and Practice*, **3**, 16-18.

Hoover, H. D. (1988). Growth expectations for low-achieving students: A reply to Yen. *Educational Measurement: Issues and Practice*, **7**(4), 21-23.

Lee, O. K., & Wright, B. D. (1992, April). *Mathematics and reading test equating*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Omar, M.H. (1996). *An Investigation into the Reasons Item Response Theory Scales Show Smaller Variability for Higher Achieving Groups.* Iowa Testing Programs Occasional Papers, Number 39.

Petersen, N. S., Kolen, M. J., & Hoover, H. D., (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221-262). Washington, DC: American Council on Education/Macmillan.

Phillips, S. E. & Clarizio, H. F. (1988a). Limitations of standard scores in individual achievement testing. *Educational Measurement: Issues and Practice,* **7**(1), 8-15.

Phillips, S. E. & Clarizio, H. F. (1988b). Conflicting growth expectations cannot both be real: A rejoinder to Yen. *Educational Measurement: Issues and Practice,* **7**(4), 18-19.

Schulz, E. M., Shen, L. S., & Wright, B. D., (1990, April). *Constructing an equal-interval scale for studying growth in reading achievement.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Seltzer, M. H., Frank, K., & Bryk, A. S., (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis,* **16**(1), 41-49.

Williams, V. S., Pommerich, M., & Thissen, D. (1995, June). *A comparison of developmental scales based on Thurstone methods and item response theory.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis, MN.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement,* **23**, 299-325.

Yen, W. M. (1988). Normative growth expectations must be realistic: A response to Phillips and Clarizio. *Educational Measurement: Issues and Practice,* **7**(4), 16-17.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics,* **17**(2), 205-218.

Notes

Work on this paper began in a research workshop sponsored by The Consortium on Chicago School Research and Chicago Public Schools, where the first author was employed. The comments of Professors Benjamin Wright, Darrell Bock, and Wendy Yen on early drafts of the manuscript are gratefully acknowledged.

Footnotes

1) If the within-grade distribution of Θ were not normal, the formulas for the median (Equation (10)) and equal-density (see text) anchor points in Figure 3 might not be precise. (Equation (10) would still be valid for symmetric distributions, where the median equals the mean.) However, it seems likely that the true median and equal density anchor points would still show substantial positive acceleration, like the anchor points in Figure 3, given the negatively accelerated trend in mean Θ. The positive acceleration is quantified by the β_2 coefficient in Equation (11). The β_2 coefficient is used in Equations (13) and (14) to approximate the impact of interpolation on conditional grade equivalent distributions, but these equations do not require the corresponding conditional Θ distributions to be normal. In mapping through integer assignment, Θ is assumed to be normally distributed within grades in order to compute weights for integer grade-equivalent values. But in this respect also, it seems unlikely that true weights corresponding to reasonable departures from normality would substantially alter the results of this study, given the degree of negative acceleration in the equal density anchor points of Figure 3.