RESEARCH
REPORT

September 2002
RR-02-18

A Model-based Approach to Measuring Cognitive Growth in Pre-reading and Reading Skills During the Kindergarten Year

Donald A. Rock
Judith M. Pollack

ETS Educational Testing Service

Research &
Development Division
Princeton, NJ 08541

A Model-based Approach to Measuring Cognitive Growth in Pre-reading and Reading Skills During the Kindergarten Year

Donald A. Rock and Judith M. Pollack

Educational Testing Service, Princeton, NJ

September 2002

Abstract

This study has two primary purposes. The first is to outline a model-based measurement approach that relies on multiple criterion-referenced points to identify different patterns of student growth in reading during kindergarten. The second purpose is to demonstrate how applied practitioners might use the model-based measurement approach to address policy questions dealing with gender and school sector differences in reading growth during kindergarten. In addition, the study shows that the more traditional approaches to measuring change with respect to these policy questions may yield uninformative if not incorrect results.

The measurement of change at the individual level is a problem that has long perplexed psychometricians (Cronbach & Furby, 1970; Harris, 1963; Rogosa, Brandt, & Zimowski, 1982; Willett, 1988; Collins & Horn, 1991; Traub, 1994; Meredith & Horn, 2001). While the problem is typically formulated in terms of low reliability of gain scores, we argue that the related but more basic problem is nonequivalence of scale units and their associated interpretations (cf. Meredith & Horn, 2001). We also argue that the use of adaptive tests may appear to increase the reliability of each test score but decrease the reliability of gain scores. That is, as one removes the sources of floor and ceiling effects through the use of adaptive measurement, all students have an equal chance to gain on the vertical scale. Depending on how adaptive is the measure, how the scale is constructed, and how even-handed is the educational treatment, one may not observe big differences in the amount of gain between individual children. It is more likely that one will see individual differences in terms of where on the scale the gain takes place. If the adaptive tests work as expected, initial status and gain should be relatively independent.

The approach taken here was to develop an item response theory (IRT) (Lord, 1980) vertical scale using an adaptive test with multiple criterion-referenced points along that vertical scale. These points model critical stages in the development of reading skills. Criterion-referenced points serve two purposes at the individual level: (1) they provide information about changes in the child's mastery or proficiency at *each* level, and (2) they provide information about *where* on the scale the child's gain is taking place. This latter piece of information about the child will be referred to as the locus of maximum gain (lcmg). The study will identify the locus of maximum gain on a hierarchical scale that is criterion-referenced to represent five critical steps in the development of early reading skills. Along with classifying children based on how much they are growing in relation to each of the five criterion-referenced points on the growth curve, we will also attempt to predict whether a child is making his/her gains on a particular critical point on the developmental scale from his/her background variables.

The study demonstrated a number of different analytic approaches to measuring cognitive growth that become available when one has a multiple criterion-referenced developmental model. We argue here that each different analytical approach brings additional insights with respect to understanding student growth. Also the methodology suggested here can be carried out on as little as two longitudinal time points.

The analysis focuses on: (1) an individual level variable (gender) and its relationship with gains, and (2) a school level variable (school sector) and its relationship with gains. In addition, this study examines the traditional approaches to measuring gain and shows where they may be uninformative in their conclusions about who gains and how much. We argue that unless one explicitly takes into consideration the location of the gain on the developmental scale, the results obtained by the traditional approaches may be misleading. The use of adaptive measurement procedures makes consideration of location of gain even more important.

Sample

The data for this study came from the Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K) sponsored by the U.S. Department of Education, National Center for Education Statistics. The ECLS-K base year sample was based on a national probability sample of approximately 20,000 children. This longitudinal study was designed to follow children's progress from kindergarten entry through spring of fifth grade. Participants in the study were tested in reading, mathematics, and general knowledge in fall and spring kindergarten and periodically thereafter. This study focused on changes in early reading skills taking place during the kindergarten year. The subsample selected to illustrate the analytic approaches to measuring change consists of ECLS-K children whose English language proficiency was judged to be sufficient to take the ECLS-K reading test and who had reading scores in fall and in spring kindergarten. In addition, the analysis sample was further restricted to children who stayed in the same school for the kindergarten year and for whom information on parent education (higher of father and/or mother) was available. The final analysis sample consisted of 13,701 children in approximately 60 private non-Catholic schools, 99 Catholic schools, and almost 700 public schools.

Test Scores

The ECLS-K reading assessment consisted of a 20-item routing test followed by three alternative second stage forms, varying in difficulty. An easy, middle, or difficult second stage form was selected to be administered to each child based on his or her performance on the routing test. The same routing and second stage tests were used on both occasions. Utilizing Bayesian procedures with PARSCALE, (Muraki & Bock, 1991) we estimated three parameters for each item pooling data from both time points, fall and spring kindergarten, and using the total item pool with separate ability priors for the two different time points. A reading scale score was defined as the estimated number correct on the total item pool (routing test items plus all second stage items). Each child's scale score was calculated by summing the probabilities of correct responses, based on IRT estimates of item parameters and the child's ability estimate. The results reported here use the scale score to illustrate traditional methods of measuring gain.

In addition to the IRT-based scale scores, probabilities of proficiency were also calculated for five skill levels that are typically taught in an ordered sequence:

- Level 1. Letter recognition: identifying upper- and lowercase letters by name
- Level 2. Beginning sounds: associating letters with sounds at the beginning of words
- Level 3. Ending sounds: associating letters with sounds at the end of words
- Level 4. Sight words: recognizing common words by sight
- Level 5. Comprehension of words in context: selecting the best word to complete a sentence

Five clusters of four items each in the routing test were used to estimate proficiency at these criterion-referenced points. Items within a cluster shared similar difficulty and content. A child was assumed to be proficient on a given cluster if he/she got any three out of four correct and was given a score of 1 for this particular item cluster or "superitem" and a score of 0 if there were two or more incorrect responses in the cluster. These superitems formed a hierarchical structure in the Piagetian sense in that the teaching sequence implies that one has to master the lower levels in the sequence before one can learn the material at the next higher level. IRT parameters for the superitems were then estimated conditional on the thetas (ability estimates) obtained from the total item pool. That is, fixing the thetas from the previous step, IRT parameters were estimated for the five superitems, treating the dichotomous pass/fail cluster score as a right/wrong item score. This parameter estimation allows one also to estimate a continuous measure of the child's probability of being proficient at each of the five critical points on the growth curve using the child's IRT ability estimate score and the parameters for each of the superitems.

Figure 1 shows the locations on the vertical scale of the difficulty levels (IRT "b" parameters) of the five superitems marking the critical points on the scale. As described above, one could then estimate the probability that a child had mastered the knowledge associated with each of the critical points on the growth curve, given the child's latent trait measure, theta. In aggregate terms, the estimated proportion of the total sample mastering letter recognition (Level 1), for example, is the average of the proficiency probability scores for that level.

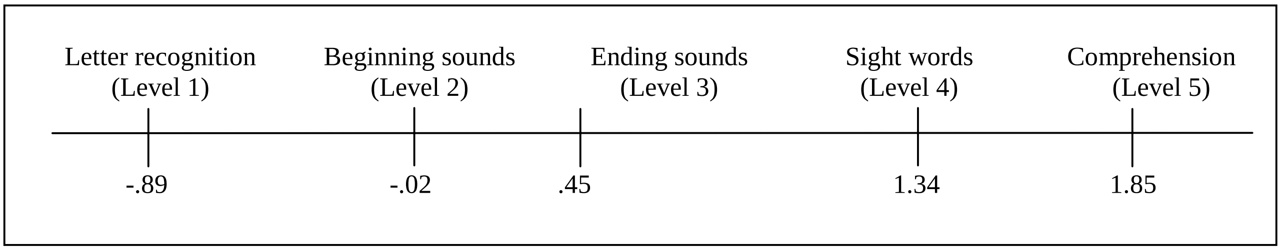

Figure 1. **Proficiency levels, theta scale.**

The basis of the reading growth model defined by the five critical points is essentially a phonics driven model, which lends itself to a hierarchical structure defined by sequential learning stages. Other models of reading development exist, e.g., the "whole language" approach, but most schools in the sample emphasized either a phonics approach alone or a more eclectic approach that includes both strategies. Empirical evidence supported the validity of the hierarchical model of proficiency levels described here in that there were very few "reversals." That is, only about 6% of the sample "passed" a higher level item cluster (three out of four correct) after "failing" at a lower level.

The critical concept of locus of maximum gain (lcmg), identifying at which of the five critical stages in development the child is making his or her maximum gain, was estimated in the following way. Differences between Time 1 and Time 2 (fall and spring kindergarten) in the probability of mastery were computed for each of the five proficiency levels. The largest difference marks the mastery level where the largest gain for a given child is taking place. This is the locus of maximum gain for that child. For example, if the largest difference in probabilities of mastery for Sheila occurs at proficiency Level 3, then we can say that Sheila is making her largest gains in the mastery of ending sounds. This simple algorithm was used to find a unique locus of maximum gain for each child. Once five mutually exclusive groups of children are identified according to the proximity of their gains to each of the five critical points on the scale, one can treat the different types of gains as qualitatively different outcome measures to be explained by background and process variables. That is, one can analyze not only how much was gained with respect to the scale scores, but also where the gains were made with respect to the superitem markers.

Method

Stata survey regression procedures (Stata Corporation, 2001) were used to: (1) model gains based on repeated measures models, with simple gains on the scale scores as the dependent variable, (2) estimate analysis of covariance models that also included the pretest scale score as a covariate for comparison with (1) above, and (3) carry out repeated measures designs that used locus of maximum gain as an additional explanatory variable in explaining a child's gain. The Stata survey regression software was used because it can account for sampling weights and cluster effects found in complex sampling designs.

Multilevel logistic regressions (Bryk & Raudenbush, 1992; Snijders & Bosker, 1999) were run to describe differences between profiles of those children who were gaining in Level 4 and Level 5 skills and profiles of children who were making their maximum gains in the levels marking the middle and lower end of the developmental scale and to show how these differences are related to variance between schools. Educational policy makers have long been concerned about the impact of where a child attends school on the child's educational growth. The so-called “school effects” can be more directly assessed within a multilevel model that explicitly defines schools as a level in the hierarchical model. Levels 4 and 5 were chosen because gains in this area of the scale occur when a child begins to read while Levels 1-3 primarily measure prereading mechanics. In the multilevel logistic regressions, the dependent variable was coded “1” if the child was making his/her maximum gain at Level 4 or 5 and “0” if the maximum gain was at Level 1 to 3. The binary dependent variable was analyzed with the school at level 2 and the child within the school at level 1. All explanatory variables were fixed, and only the intercepts were considered random. These multilevel logistic regressions speak directly to the question of whether children who were changing at or above this critical point on the developmental scale came from different backgrounds and attended different types of schools than those children changing below this point (i.e., growing in their prereading mechanical skills). Many researchers do not have access to statistical software packages that can estimate the appropriate standard errors for complex sample designs. However, multilevel software packages are becoming more readily available, and they can correct the standard errors for the clustering effects present in many hierarchical sample designs. The multilevel logistic regressions were estimated using quasi-likelihood estimators available in the MlwiN software (Goldstein et al., 1998). Examples of multilevel approaches to measuring change within the present model-based criterion-referenced framework will also be presented here.

Results

Table 1 presents the fall and spring kindergarten reading means, standard deviations, and correlations in the scale score metric for the subsample used in this analysis. The standard deviations increase from Time 1 (fall kindergarten) to Time 2 (spring kindergarten). The increase in standard deviations suggests the potential of observing a "fan spread effect" (Campbell & Erlebacher, 1970). The correlation of .03 between initial status and gain suggests little or no linear relationship between initial status and amount of gain in the total scale score metric. It appears that the adaptive test worked as expected, minimizing floor and ceiling effects. This low correlation between initial status and gain suggests that the standard analysis of covariance approach that controls for initial status and the repeated measures approach that analyzes the simple difference scores will yield very similar results.

Table 1

Fall and Spring Kindergarten Reading Scale Scores—Means, Standard Deviations, and Correlations

	Time 1	Time 2	Gain
Mean	22.65	32.56	10.07
Standard deviation	8.55	10.81	6.32
Correlations:			
Time 1	1.00		
Time 2	.81	1.00	
Gain	.03	.61	1.00

The analysis of total scale scores by gender and school sector reported in the next sections shows nearly identical average gains from fall to spring kindergarten of about 10 scale score points for all subgroups, even though the mean scores for the subgroups are quite different. Individual and group differences in the amount of gain given a fairly standard treatment (the year of kindergarten schooling) can be relatively trivial compared to individual and group differences in the areas where the gains take place. Thus analysis of total scale score without explicitly taking into consideration where the gain takes place tells only part of the story.

Gender and Locus of Maximum Gain

Figure 2 shows a plot of fall to spring kindergarten reading gain by gender in the total scale score metric, adjusted for age at first testing, time lapse between testing, and highest level of parent education. Figure 2 indicates that girls were more advanced in reading when they entered kindergarten and had increased their advantage by the time they were retested. In terms of the classical repeated measure analysis, there is a significant interaction between gender and time of testing, indicating differential gain (F=27.66; df=1, 96; p=.00). The degrees of freedom in the denominator reflect the number of primary sampling units in the complex sample design minus the number of predictors. All the Stata-regression-based F tests reported here are based on adjusted Wald tests (Korn & Graubard, 1990). Using the classical ANCOVA with the pretest, the highest level of parent education, and the two age-related variables as covariates yield adjusted spring reading means that significantly favor girls (F=29.48; df=1, 95; p=. 00). As expected, the ANCOVA and repeated measures yielded similar F tests since the gain scores were essentially uncorrelated with initial status.

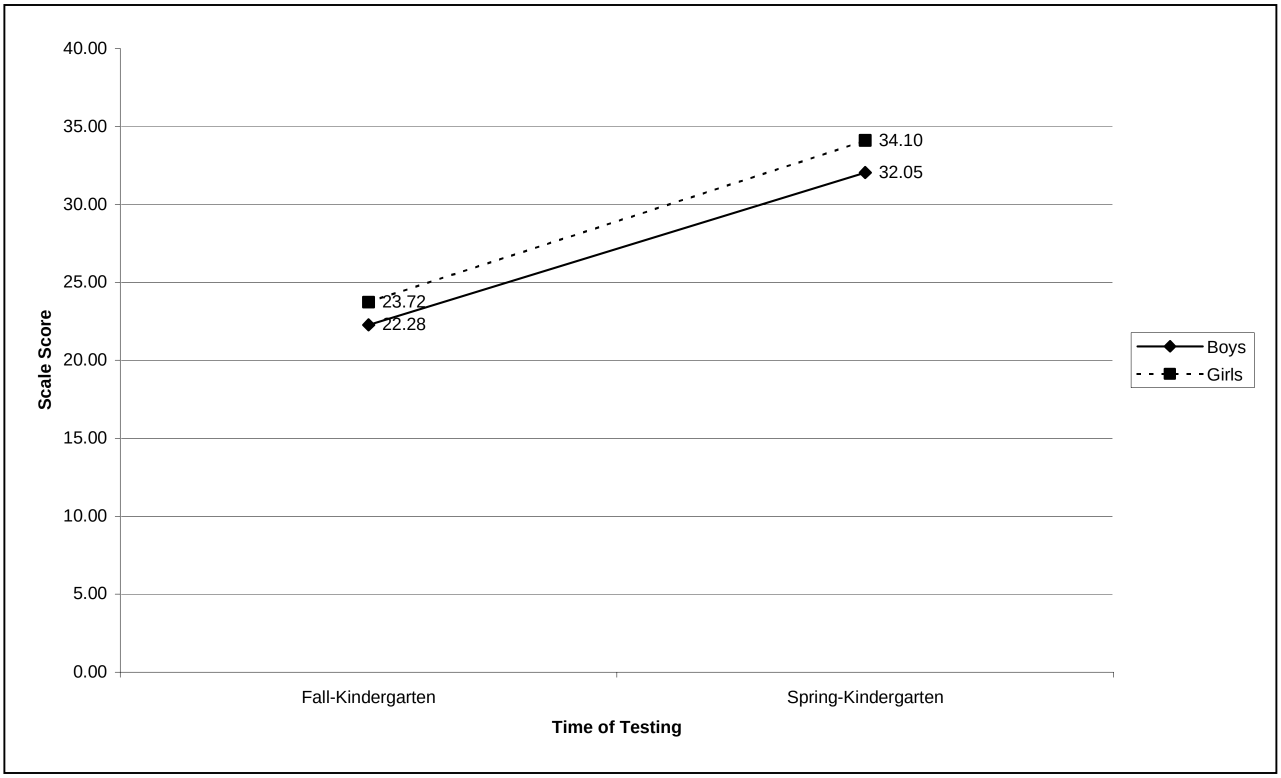

Figure 2. **Total scale score metric of fall to spring reading gain by gender, adjusted for age, time lapse between testing, and parent education.**

Figure 3, a clustered histogram showing the locus of maximum gain by gender, indicates that slightly over one third of the boys are making their maximum gains in letter recognition, while about a quarter of the girls do the same. Girls are more likely than boys to be making their maximum gains in the middle level proficiencies, beginning and ending sounds and sight words. Only a very small percentage of both boys and girls make their greatest gains at the highest level (comprehension of words in context) during the kindergarten year.

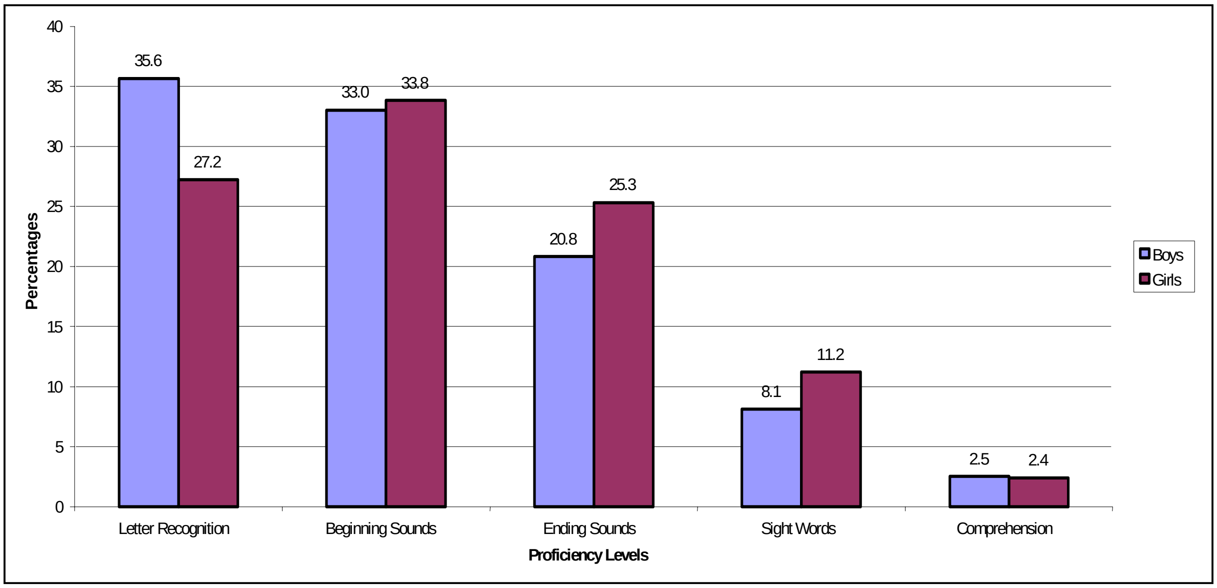

Figure 3. **Locus of maximum gain by gender, fall to spring kindergarten.**

The next step in the analysis was to run the multilevel logistic regression described in the Method section, with children making their maximum gains in beginning reading (Levels 4 and 5) coded "1" and children making their maximum gains in prereading mechanics (Levels 1–3) coded "0." In the first set of these multilevel regressions, gender was the explanatory variable of specific interest, with highest parent education, age at first testing, and time lapse between testing used as covariates. Table 2 presents the gender logistic partial regression weights and their associated odds ratios and tests of significance. All of the explanatory variables were considered fixed and only the intercept was considered random.

Table 2

Gender Multilevel Binomial Analysis of Gain at Levels 4-5 Versus Gain at Levels 1-3

Explanatory variables (1)	Logistic regression weights (2)	Odds ratios for main effects of gender equation (3)	"*t*" statistics (*P*-value) (4)	Marginal reduction in between school variance (5)
Age first testing	.07	1.07	11.50 (.00)	0.00%
Time lapse	.04	1.04	.69 (.50)	
Parent education				
High school graduate	1.27	3.56	5.90 (.00)	
Some college	1.70	5.46	7.93 (.00)	32.16%
College graduate	2.47	11.80	11.47 (.00)	
Postgraduate	2.92	18.46	13.37 (.00)	
Girls	.35	1.42	6.62 (.00)	0.00%

The last column of Table 2 presents the marginal reduction in the between school variance (compared to an intercept-only null model) that was due to: (1) maturation as measured by age at first testing and the time lapse until the second testing, (2) the block of dummy variables contrasting various parent educational levels with the base level (less than high school), and (3) the variable of interest—gender. For example, the block of parent education contrasts reduces the between-school variance by about 32% from that present in the null model (i.e., the intercept-only model). Clearly, there are large differences between schools with respect to parent education, and those differences are related to where on the scale the children are making their gains.

Columns 2-4, the logistic regression weights, odds ratios, and *t* statistics, are based on the full model—that is, every explanatory variable is entered into the model. Inspection of column 3, the odds ratios, indicates that girls are almost one-and-one-half times (1.42) more likely than boys to be making their maximum gains at the two highest levels of the reading developmental scale when age differences and parent education are controlled. The odds ratios associated with children of college-educated parents show a disproportionately larger increase compared to the other contrasts. There was little or no between-school variance associated with gender variance in gains as shown by the lack of marginal reduction in between school variance associated with the girls dummy code. This suggests that the pattern of gender differences in gains at Levels 4 and 5 versus Levels 1–3 varies little across schools when controlling for parent education and age differences.

The above analyses were primarily concerned with where on the developmental scale the children are making their maximum gains. Figure 4 contrasts the genders on the *amount* of their scale score gains by groups defined by their locus of maximum gain. The points in Figure 4 represent the adjusted cell means from a two-way, repeated measures design with gender and locus of maximum gain as the design factors and scale score gain as the dependent variable. Cell means were score gains adjusted for all covariates: age at first testing, time lapse between testing, and highest level of parent education. Table 3 presents summary statistics for this design. The part correlations associated with each factor and their interaction are also shown to illustrate the relatively small effects associated with differential gain.

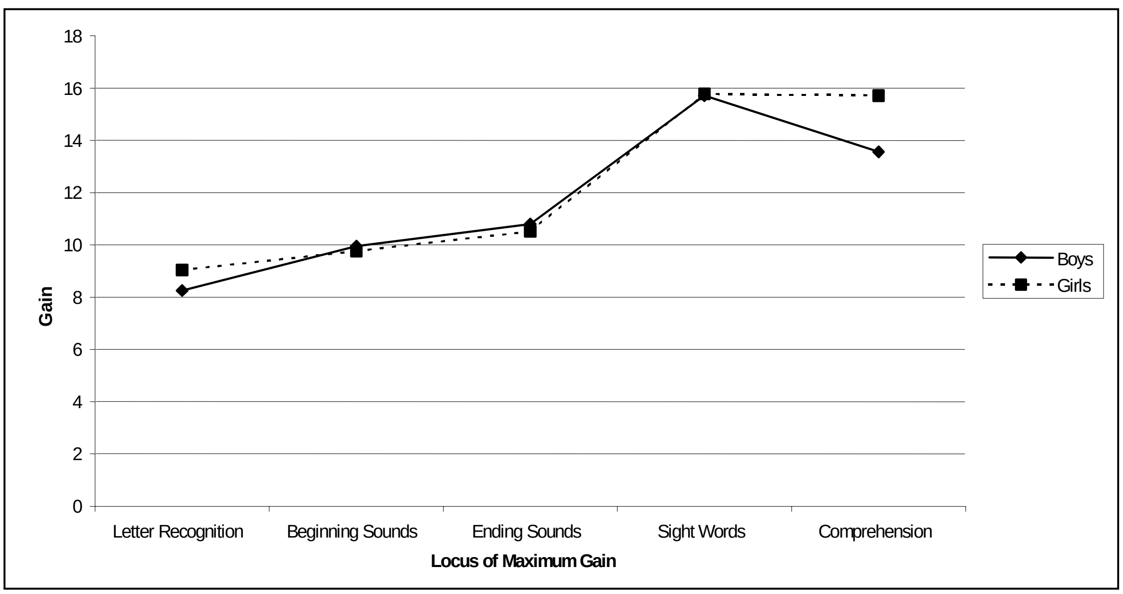

Figure 4. **Adjusted mean gain in total scale score by gender and locus of maximum gain.**

Table 3

Summary of the Repeated Measures Analysis of Total Scale Score Gains by Gender and Locus of Maximum Gain

Design factors	Degrees of freedom	*F* based on adjusted Wald tests	"*F*" statistic (*P*-value)	Part correlation
Gender	1, 99	18.81	.00	.04
Locus of maximum gain	4, 96	104.05	.00	.05
Locus of gain x gender	4, 96	6.59	.00	.05

Note: Controlling for age at first testing, time lapse between testing, and parent education.

Figure 4 indicates that while girls and boys are about equally represented at the highest level (about 2.5% of either gender group), girls at this level appear to be making larger gains than boys. The adjusted Wald test for this simple main effect was significant (F=5.57; df=1, 99; p=.02). Conversely, considerably more boys than girls are at Level 1, (see Figure 3), but those girls who are making their maximum gains at this level are making slightly greater gains than the boys. The test of their differential gain is statistically significant (F=4.95; df=1, 99; p=.00).

School Sector and Locus of Maximum Gain

Figure 5 presents the fall to spring total scale score gains by school sector adjusted for age, time lapse between testing, and highest level of parent education.

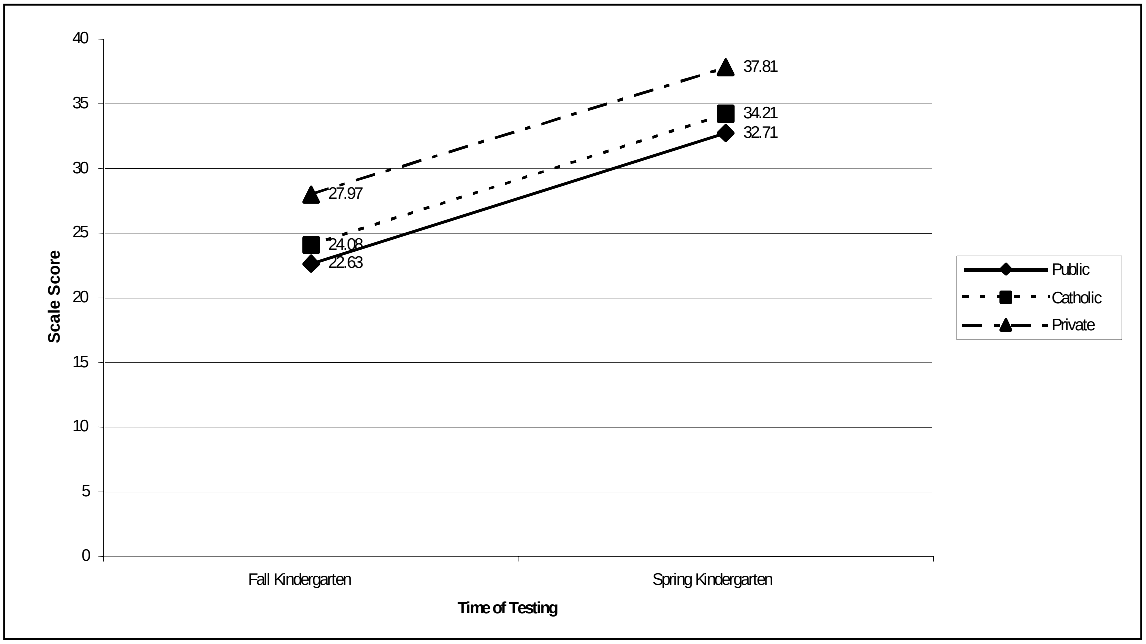

Figure 5. **Total scale score metric of fall to spring reading gain by school sector, adjusted for age, time lapse between testing, and parent education.**

Figure 5 suggests that in terms of the total scale score metric, schools from the different sectors have different initial statuses with respect to their children's developmental level, and they appear to maintain the same relative positions on retesting in the spring. In terms of the traditional repeated measures approach, the school sector by trial interaction (F=.25, df=2, 98, p=.78) was clearly not significant. An ANCOVA analysis with the fall kindergarten test scores, age, time lapse, and highest level of parent education as covariates yields almost exactly the same results (F= .26, df=2, 98, p=.77). That is, there is no difference in the amount of gain for children from the different school sectors in terms of the total scale score metric. This result changes when we explicitly take into consideration where the children are making their maximum gains.

Figure 6 presents a histogram detailing where on the scale the children from different school sectors were making their maximum gains. It shows quite different patterns with respect to where the gains were taking place.

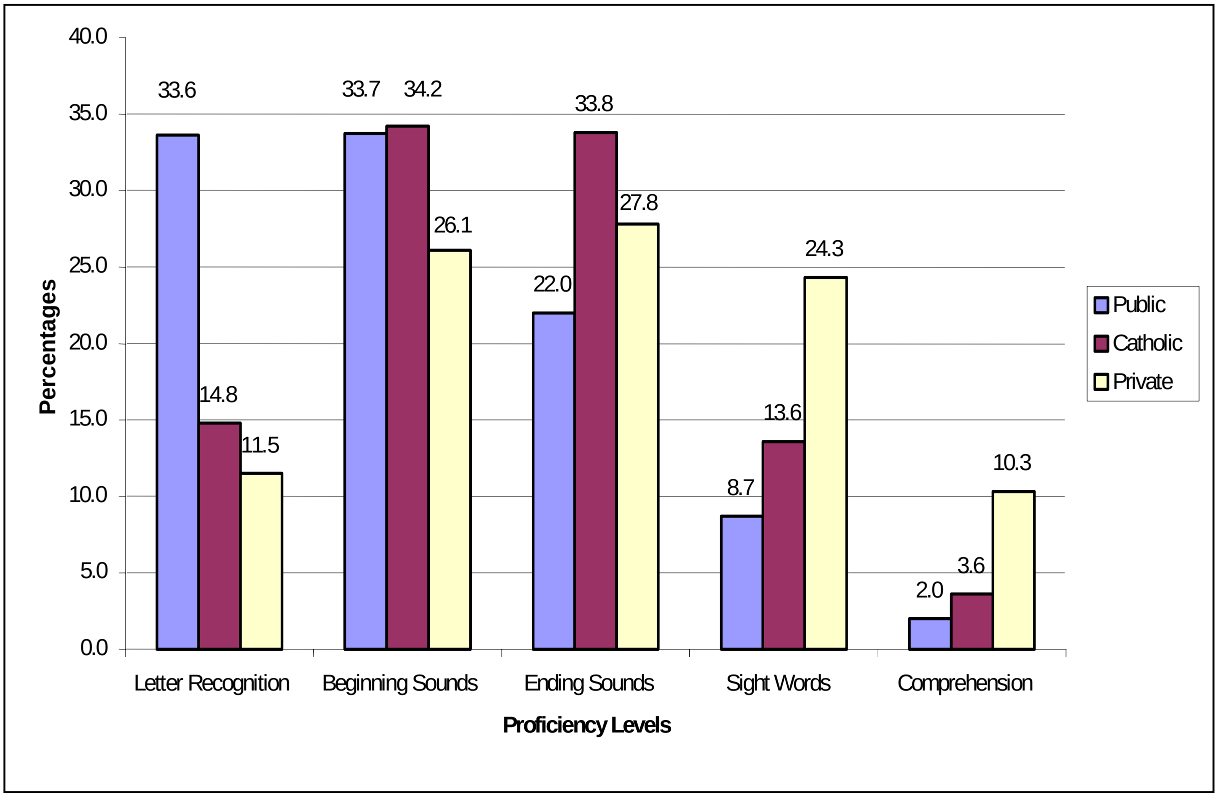

Figure 6. **Locus of maximum gain by school type, fall to spring kindergarten.**

Figure 6 indicates that a full one third of the public school children were making their maximum gains at the lowest level skill (letter recognition), while only 11% of the private non-Catholic school children were making their maximum gains at this level. Conversely, one third of the private non-Catholic school children compared to about 11% of the public school children were making their maximum gains at the early reading skills—sight words and comprehension of words in context (Levels 4 and 5). Contrasts between the Catholic and public school children with respect to the locus of maximum gain indicates that two thirds of the Catholic school children were making their maximum gains in beginning sounds and ending sounds (Levels 2 and 3), while a similar proportion of the public school children were making their maximum gains in letter recognition and beginning sounds (Levels 1 and 2).

Table 4 presents the multilevel logistic partial regression weights relating the school sector explanatory variables (controlling for age at time of first testing, time lapse between testing, and parent education) to the dichotomous outcome of whether the child was making his/her maximum gains at Levels 4 or 5 (early reading) versus the lower three levels (prereading mechanics).

Table 4

School Sector Multilevel Binomial Analysis of Gain at Levels 4-5 versus Gain at Levels 1-3

Explanatory variables (1)	Logistic regression weights (2)	Odds ratios for main effects of school sector equation (3)	"*t*" statistics (*P*-value) (4)	Marginal reduction in between school variance (5)
Age first testing	.07	1.07	11.33 (.00)	0.00%
Time lapse	.05	1.05	.96 (.32)	
Parents education				
High school graduate	1.27	3.56	5.79 (.00)	
Some college	1.68	5.38	7.68 (.00)	32.16%
College graduate	2.43	11.31	11.08 (.00)	
Postgraduate	2.84	17.10	12.73 (.00)	
Catholic	.14	1.15	1.11 (.26)	6.62%
Private non-Catholic	1.07	2.91	7.02 (.00)	

In this multilevel logistic regression analysis, the public schools are the base or contrast group. Inspection of the odds ratios in column 3 of Table 4 indicates that the private non-Catholic school children are almost three times as likely as public school children to be making their maximum gains in the higher level, beginning reading skills (Levels 4 and 5). There is no significant difference between the Catholic school children and the public school children with respect to this dichotomous criterion.

Column 5 shows that the school sector variables reduced the between school variance by more than 6%. However, the major part of the explainable between-school variance in cognitive growth in beginning reading skills (Levels 4 and 5) is associated with parent education and more specifically, college-educated and postgraduate parents. This indicates that regardless of school sector, children from homes with college educated parents are more likely to enter kindergarten already knowing their prereading mechanics and ready to progress in early reading skills.

Figure 7 shows the amount of gain in total scale score points taking place for groups defined by their locus of maximum gain crossed with school sector. This figure shows graphically the results of a school sector by locus of maximum gain repeated measures design.

The repeated measures scores were adjusted for all covariates: age at first testing, time lapse between testing, and highest level of parent education. Inspection of Figure 7 suggests that when locus of maximum gain is controlled, public school and Catholic school children are consistently gaining as much or more than the children in the private non-Catholic schools. In fact, post-hoc–adjusted Wald tests show that overall the private non-Catholic school children are gaining less than the public school children (F=21.90; df=1, 99; p=.00) and also less than the Catholic school children (F=11.69; df= 1, 99; p=.00). There are no differences between public school children and Catholic school children in overall gains. The overall gains in favor of the public school children compared to the private non-Catholic school children are due to relatively large differential gains at Level 1, letter recognition (F=21.90; df= 1, 99; p=.00) and to a lesser extent Level 2, beginning sounds (F=4.66; df=1, 99; p=.03).

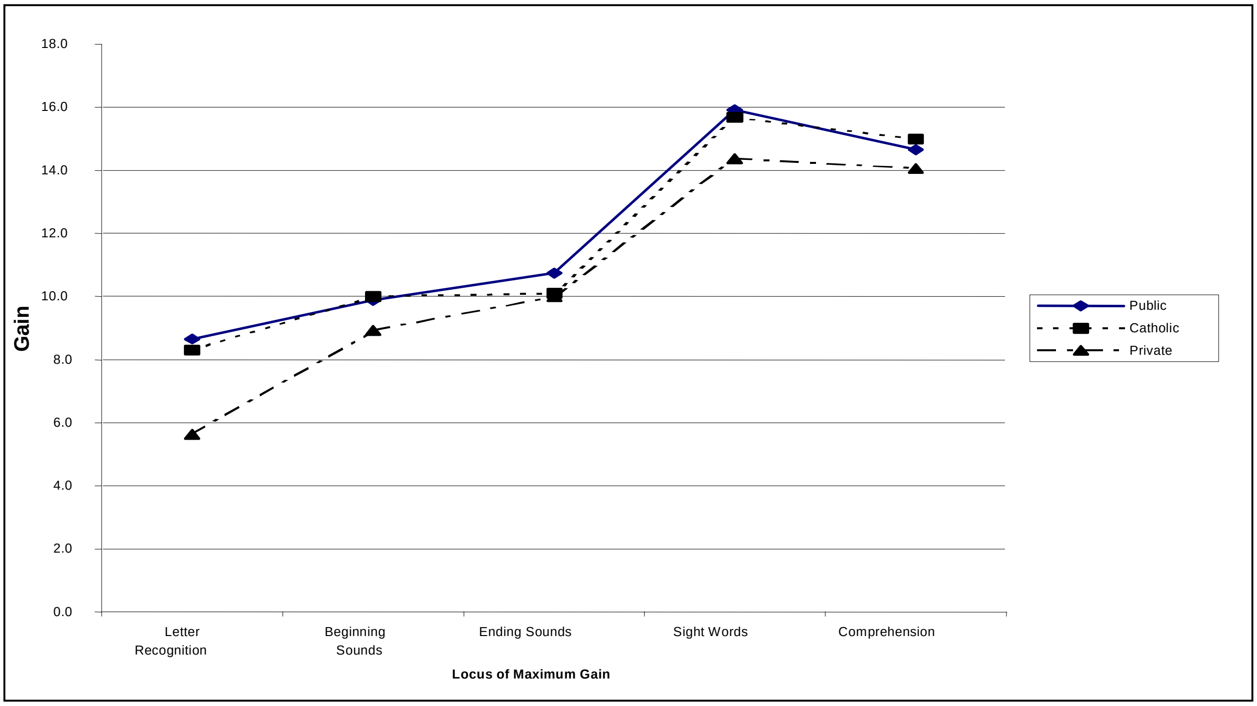

Figure 7. **Adjusted mean gain in total scale score by school sector and locus of maximum gain**

Table 5 summarizes the repeated measures of the gain scores for school sector crossed with locus of maximum gain. While all the main effects and interaction are statistically significant, the main effect of the locus of maximum gain overwhelms the other factors as indicated by both the adjusted Wald test and the part correlations.

Table 5

Summary of the Repeated Measures Analysis of Total Scale Score Gains by School Sector and Locus of Maximum Gain

Design factors	Degrees of freedom	*F* based on adjusted Wald tests	"F" statistic (*P*-value)	Part correlation
School sector	2, 98	11.69	.00	.03
Locus of maximum gain	4, 96	172.90	.00	.30
Locus of gain x school sector	8, 92	2.29	.03	.03

Note. Controlling for age at first testing, time lapse between testing, and parent education.

It appears that children entering kindergarten with only Level 1 or Level 2 skills may gain more at the public or Catholic schools. In fact, the results graphed in Figure 7, which are consistent with the significant simple main effects discussed above, suggest that the public and Catholic school kindergartens seem to be doing creditable jobs at all levels of the developmental scale and not just the lower levels.

An Alternative Measure of Overall Gain

As pointed out above, the complex patterns that can occur with respect to gain scores are not always properly summarized in a single overall measure of gain, whether they are raw gains from repeated measures or residualized gains from ANCOVA with the pretest as a covariate. The use of adaptive tests may make this even more complicated. While we have argued that no single summary gain score will give entirely satisfactory results, there are times when the convenience of having a single score may compensate for some of the lost information. Also, in many instances, the measurement instrument may not be constructed to criterion-referenced critical points on a growth curve. The percent of maximum possible gain is suggested as an alternative single summary score that takes into consideration where the gain is taking place on the developmental scale. This score implicitly assumes, however, that gains at the upper end of the scale are more important than gains at the lower levels. The logic for this is as follows: At a given point in time in a developmental process such as learning to read, those children who are making their gains at the upper end of the developmental scale, which typically reflect more complex tasks, will be better positioned for further advancement in reading skills. In addition, as they become more skilled in reading comprehension, they will be able to use reading as a tool in mastering other school related skills. Equation 1 below estimates the percent of maximum possible gain as follows:

$$\hat{Y}_{gi} = \left[\left(y_{i2} - y_{i1} \right) / \left(y_{\max} - y_{i1} \right) \right] x100. \qquad (1)$$

where: $\hat{Y}_{gi}$ = percent of maximum gain for individual i

y_{i2} = total scale score at time 2 for individual i

y_{i1} = total scale score at Time 1 for individual i

$y_{\max}$ = maximum possible total scale score on the item pool.

The percent of maximum possible gain as defined in Equation 1 also has the potential for helping to minimize the impact of ceiling effects if they should occur. Percent of maximum gain can be viewed as a variation on the POMP score suggested by Cohen, Cohen, Aiken, and West (1999) for measuring status at a single point in time.

Table 6 compares multilevel results for raw gains with the percentage of maximum gain. Inspection of Table 6 suggests that the block of parent education variables becomes considerably more important when the outcome is percentage of maximum gain. This is not surprising since the percentage of maximum gain includes a component related to where on the scale the gain takes place. More importantly, the signs of the school sector variables have gone from negative to positive and the private non-Catholic regression weight comes close to significance. These findings come closer to what is known about the performance of the children in the private non-Catholic schools. That is, they are disproportionately represented in the group making their maximum gains at the upper end of the scale. When data from the first grade follow-up of ECLS-K is analyzed, this measure of gain may be more effective than other scores in summarizing what is really happening. The present longitudinal scale is fixed through the spring of first grade, and thus more children will be approaching the upper end of the scale at that time.

Table 6

Multilevel Analysis Comparison of Total Scale Score Gain Outcomes With Percentage of Maximum Gain Outcomes

Explanatory variables	Scale score			Percentage of maximum gain		
	Regression coefficient (SE)	"Z" statistic (*P* value)	Marginal reduction in between school variance	Regression coefficient (SE)	"Z" statistic (*P*-value)	Marginal reduction in between school variance
Age first testing	-.01 (.01)	.28		.06 (.02)	.00	
Time lapse	1.12 (.11)	.00	17.45%	1.59 (.17)	.00	13.52%
High school graduate	.78 (.20)	.00		1.65 (.33)	.00	
Some college	1.29 (.21)	.00	.68%	2.71 (.34)	.00	7.32%
College graduate	1.57 (.23)	.00		3.90 (.36)	.00	
Postgraduate	2.00 (.26)	.00		5.35 (.41)	.00	
Catholic	-.13 (.33)	.70		.24 (.51)	.62	
Private non-Catholic	-.40 (.43)	.35	0.00%	1.04 (.66)	.11	.54%

Discussion

Children attending public, Catholic, and private non-Catholic schools appear to be making equivalent gains when the traditional methodology is applied using the total scale score metric. However, the pattern with respect to where on the scale they are making their gains is quite different. Public school children and to a lesser extent Catholic school children are both overrepresented proportionally and making the biggest gains at the lower end of the developmental scale. Conversely, the children attending private non-Catholic schools tend to be overrepresented among those gaining at the upper level reading skills, yet they are not gaining more in scale score points than their public and Catholic school counterparts who are gaining at these levels. The failure of traditional methods of analyzing gains (repeated measures and

ANCOVA) to take these distributional differences into account leads to conclusions that there are no school sector differences with respect to gains in reading in kindergarten. It is only when we explicitly model where the changes are taking place and then compare the school sectors within these locations of change that we find systematic differences between the school sectors. If the primary task of public and Catholic school kindergartens is to raise the reading skills of all children regardless of how deficient they may be at entry, then they are doing quite a credible job. These data do not demonstrate an advantage for private non-Catholic schools over public and Catholic schools in raising children's reading scores when initial status is controlled, even for children who started at the highest achievement levels.

In the case of gender differences, both the traditional approaches and the approaches that model the locus of maximum gain yield similar results. One of the reasons for this agreement among methods is the fact that girls were not only overrepresented at the upper end of the developmental scale (where gains tended to be larger) but also had equivalent or greater gains at each of the other levels.

Clearly, many of the children entering the different school sectors are beginning kindergarten with quite different reading skills. However, children who leave kindergarten with beginning reading comprehension skills (Levels 4 and 5) will not only have the advantage of possessing higher level reading skills but will have a comprehension tool that they can use to enhance their learning in other areas and possibly widen the gap in later years between themselves and their counterparts who are growing in the lower level reading skills. It is for this reason that children in the private non-Catholic schools may have the potential in future follow-up assessments to increase the gap between themselves and their counterparts in the public and Catholic schools.

Conclusions

The methodology used in this analysis utilized adaptive tests with multiple criterion-referenced points that mark critical points in the early reading developmental process. Emphasis was placed on where on the vertical scale the gain was taking place as well as the amount of gain. The results show the following:

- Traditional approaches to measuring gain found no differences between school sectors. However, when the locus of maximum gain was explicitly controlled, children at both public and Catholic schools showed significantly greater gains than did their counterparts at private non-Catholic schools. While parent education is highly related to where children's gains are being made, and thus, of course, directly to reading skills at input, controlling for pretest scores and parent education may be misleading. The reasons for this are tied up in the distributional differences with respect to where the gains are taking place as well as a nonlinear relationship between amount of gain and where on the scale that gain is taking place. In summary, it would seem that the public and Catholic school kindergartens are doing quite a good job given the wide range of reading skills among children entering kindergarten. On average, the private non-Catholic school children enter kindergarten with more advanced reading skills than their counterparts in the other school sectors. Thus they are overrepresented with respect to growing the skills measured at the upper end of the scale dealing with beginning reading, and because of their head start they may have the potential of widening the gap in the future.

- Girls on average begin kindergarten with better prereading skills than do boys.
- On the whole, girls gain more than boys in the total scale score metric, and this finding is independent of the analytic method used.
- Boys and girls differ on where on the scale they make their maximum gains. Boys are almost twice as likely as girls to be making their maximum gains in the lowest level prereading skill, letter recognition.
- However, among girls and boys who are making their maximum gains at Level 1 (letter recognition), girls gain more than boys during the kindergarten year.
- Girls are more likely than boys to be making their maximum gains in the areas of the scale dealing with ending sounds (Level 3) and sight words (Level 4).
- Girls and boys have about equal representation among children who are gaining at the upper end of the scale defined by Level 5 skills (comprehension of words in context).
- Among girls and boys who are making their maximum gains at Level 5 (comprehension of words in context), girls gain more than boys during the kindergarten year.
- Parent education is closely related to where the gains are taking place on the developmental scale but has little relation to the amount of gain once the locus of maximum gain is entered into the model.

- On average, children in public schools had the lowest reading skills on entry to kindergarten, followed by children attending Catholic schools, with children entering private non-Catholic schools having the highest reading skills.
- Children attending public schools are much more likely than children attending Catholic or private non-Catholic schools to be gaining on Level 1 tasks (letter recognition).
- Children attending Catholic schools are much more likely to be making their maximum gains in Level 3 skills (ending sounds) than children attending either public or private non-Catholic schools.
- Children attending private non-Catholic schools are much more likely than children at public or Catholic schools to be making their gains in Level 4 (sight words) and Level 5 (comprehension of words in context). This differential in favor of private non-Catholic schools was particularly large for the Level 5 tasks.
- While proportionately more private non-Catholic school children are making their maximum gains at Levels 4 and 5, they are not gaining more in total scale score points than their counterparts from the public and Catholic schools who also are making their maximum gains at those levels.

References

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models, applications and data analysis methods.* Newbury Park, CA: Sage Publications.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory education: A national debate.* New York: Brunner/Hazel.

Cohen, P., Cohen, J., Aiken, L., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*(3), 315-346.

Collins, L. M., & Horn, J. L. (Eds.) (1991). *Best methods for the analysis of change.* Washington, DC: American Psychological Association.

Cronbach, L .J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin, 74,* 68-80.

Goldstein, H., Rashback, J., Plewis, I., Draper, D., Browne, W., Yang, M., et al. (1998). *A user's guide to MLwinN.* London: Multilevel Models Project, Institute of Education, University of London.

Harris, C. W. (1963). *Problems in measuring change*. Madison: University of Wisconsin Press.

Korn, E. L., & Graubard, B. I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *The American Statistician, 44,* 270-276.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Publishers.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change.* Washington, DC: American Psychological Association.

Muraki E. J., & Bock, R. D. (1991). PARSCALE: Parameter scaling of rating data [Computer software]. Chicago, IL: Scientific Software, Inc.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726-748.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage Publications.

Stata Corporation. (2001). *Stata 7 user's guide*. College Station, TX: Stata Press.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage Publications.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345-422). Washington, DC: American Educational Research Association.