Research Report

A Note on Gain Scores and Their Interpretation in Developmental Models Designed to Measure Change in the Early School Years

Donald A. Rock

Research & Development

March 2007
RR-07-08

A Note on Gain Scores and Their Interpretation in Developmental Models Designed to Measure Change in the Early School Years

Donald A. Rock
ETS, Princeton, NJ

March 2007

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This paper presents a strategy for measuring cognitive gains in reading during the early school years. It is argued that accurate estimates of gain scores and their appropriate interpretation requires the use of adaptive tests with multiple criterion referenced points that mark learning milestones. It is further argued that two different measures of gain are necessary because of the nonequivalence of scale score points. Empirical results are presented in support of the use of adaptive tests with multiple criteria referenced scale points for the measurement of gain.

Key words: Adaptive tests, criterion referenced, gain scores, reading, locus of maximum gain, learning milestones

Introduction

This short paper discusses the uncritical use of gain scores and their interpretation when estimating their relationships with process and background variables. To support this position, the author argues for supplementary and alternative approaches using reading data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998 (ECLS-K). The ECLS-K, sponsored by the National Center for Education Statistics (NCES) of the U.S. Department of Education, assessed a representative sample of approximately 22,000 of the nation's children entering kindergarten in the fall of 1988. These children were then followed up and re-tested on selected occasions through the fifth grade. A large nationally representative sample of children entering kindergarten will yield a very heterogeneous distribution of cognitive scores whether one is talking about early language and literacy or number skills. The resulting range of ability argues for an adaptive test, which has the property of maintaining a relatively constant level of measurement precision throughout the range of ability by tailoring item difficulty to the examinee's ability level. The implementation of adaptive measures becomes even more critical if the purpose is to measure change and relate that change to teachers, parents, and school process variables. It is argued here that the estimation of gain scores alone, whether raw or adjusted, will lead to biased estimates of children's growth trajectories and/or their relationships with process variables unless:

1. The scores are estimated using adaptive testing procedures and
2. The test score scale has multiple criterion referenced points that ideally mark crucial developmental milestones.

Gain scores that *are not* based on an adaptive testing approach are likely to give erroneous results because of floor and ceiling effects. Tests that are too hard or too easy will be unreliable in both tails of the score distribution and are likely to underestimate the amount of gains for students in both tails. Gain score analysis of classrooms populated by a large proportion of children at either end of the ability distribution at the fall or spring assessment would likely yield incorrect estimates of both student trajectories as well as the relationships of gain with teacher, parent, or school process variables.

While adaptive tests have the potential of estimating gain scores with virtually equal precision all along the test score scale, they do not by themselves provide information on what operational skills a given child is making progress towards proficiency on. Depending on where

on the test score scale the child is making gains, we can infer that the child is making progress on the skills being measured by items on the scale in the area where the change is taking place. Ideally when the adaptive item pool is developed, marker items—items that describe critical steps in the accumulation of language skills or mathematical knowledge—can be identified and subsequently located at various ascending points on the test score scale. The scale can then be said to have multiple criterion referenced points marking a hierarchy of proficiencies. Item response theory (IRT) scaling (Lord, 1980; Embretson & Reise, 2000) of items makes this possible since the item difficulties are on the same scale as the latent ability (theta) score. They can then mark critical points on the score scale indicating learning milestones. In addition to having an estimate of the amount of the value added based on the size of the gain score, the policymaker also will have information on which skills a given child was showing improvement. This latter concept (i.e., the notion of not only measuring how much a child gains but also providing information about where on the scale the child is making maximum gains) will be referred to as a child's *locus of maximum gains.* More specifically the locus of maximum gain for an individual is the proficiency level where the child is making the largest relative gain. The measurement of gain is only complete when we know not only how much a child gains but where on the score scale (i.e., at which proficiency level the child is making the largest relative gain). The ECLS-K battery was designed to do both.

The fact that we now have at our disposal powerful techniques for estimating individual growth trajectories and their correlates—namely, the hierarchical modeling techniques made operational in such software as MLwin (Goldstein, 1998) and HLM (Raudenbush et al., 2001) — doesn't negate the necessity of furnishing unbiased and interpretable estimates of the growth variable as input to these powerful programs.

The question then becomes what type of adaptive test would be most efficient from both an operational and technical perspective for measuring status and gains for young children? There are many variations possible in both delivery systems and the extent of adaptivity. Computer-adaptive tests (CAT) vary from completely computer-driven adaptive tests (CCAT) to computer assisted adaptive tests (CAAT). CCAT typically present the items on the computer screen and the respondee answers by using the touch screen, mouse, or inputting the answer using the keyboard. This obviously would not be appropriate for children entering kindergarten. The extent of adaptability also can vary. Completely adaptive procedures use a real time

computer algorithm to select the next item that is most appropriate from a precision standpoint for a given child depending on this individual's performance on previous items. Psychometric theory suggests that maximum measurement precision is achieved when a child is given an item whose difficulty level is nearly the same as the child's overall ability level. Since IRT models allow item difficulties and latent trait scores to be on the same scale metric, adaptive tests can score a child in real time and select the next item consistent with this individual's ability level. In theory the computer will stop administering items when the latent trait score reaches a pre-determined level of precision. However, in the real world with real world item pools often a nontrivial proportion of children never reach the desired precision level and a stopping rule has to be applied.

A procedure that is less than completely adaptive called the two stage adaptive test procedure (Cleary, Linn, & Rock, 1968; Lord, 1980) was considered and deemed to be appropriate for the ages and abilities present in kindergarten and subsequent early grades. This alternative adaptive scheme was composed of a two-stage computer *assisted* adaptive test rather than a completely computer-driven administration. This particular computer assisted adaptive approach is particularly appropriate for one-on-one testing situations where the children cannot be assumed to be proficient on the computer. In this type of administration the child is administered a short test consisting of items covering a broad range of difficulty called a routing test (i.e., the first stage) and then the examinee is sent to one of three second stage tests depending on how well the given child did on the routing test. If a child received a low score on the routing test this child will be sent to an easier second stage test. Children that receive high scores on the routing test are sent to a harder second stage test, and so on. Thus the second stage test is tailored to the ability of the child. The job of the test administrator is to individually administer the items in the routing test and then input the child's answers which the computer scores and then it in turn tells the test administrator which second stage test should be administered. A well constructed two-stage adaptive test will only be marginally less efficient (in testing time and precision of measurement) than a completely adaptive test. The completely adaptive test achieves this marginal superiority by having a much larger item bank with items at every possible difficulty level enabling the branching algorithm to pick the next item to administer to more precisely zero in on an examinee's skill level.

The advantages of the two-stage CAAT are many and they include:

1. One can measure each child no matter where they are in the ability distribution with good precision (i.e., with small standard errors of measurement). This results from matching the item difficulties to the child's ability level. This property of invariance of measurement precision across the score scale is absolutely essential for measuring gains since floor and ceiling effects will be minimized.
2. It is likely to improve the motivation and interest of the examinee since individuals are given items appropriate to their respective ability level.
3. One can achieve the same precision as a standard test with about a third to half of the testing time.
4. Longitudinal studies such as the ECLS must be able to put scores from the planned follow-ups on the same scale so that the user can validly measure gains over time and grade. This is known as the vertical scaling problem.

The larger the gap in time between test administrations the more growth is likely to take place and the greater the vertical scaling problem. Adaptive tests minimize this problem because of their range of item difficulties and thus they are more likely to provide the necessary linking items that can overlap the highest ability level of say spring first graders with the lower ability levels of spring third graders. Since adaptive tests can reliably measure the upper end of the ability of spring first graders (no ceiling effects) as well as the lower end of abilities of third graders (no floor effects) we can have reliably measured overlapping scores from both ability distributions based on common linking items. This is precisely the situation in ECLS where second graders were not tested so the vertical scale had to span from spring first grade to spring third grade.

IRT-based methods such as two-stage adaptive tests can sometimes lead to score interpretation problems however, because each child is taking different items depending on their skill level. Content or curriculum specialists often perceive a problem here since every child is not receiving the appropriate allocation of each of the content areas/process areas represented in the total item pool. This is a somewhat specious complaint in the sense that the IRT scoring system creates a *weighted* score that takes into consideration the item parameters in the weighting while those same item parameters were estimated on the total item pool, which was indeed designed to match the test specifications. A second problem has to do with the notion that

the item pool must be grade- or age-level appropriate. By there very nature adaptive tests must include items that many curriculum specialists might judge to be *grade-level inappropriate.* An adaptive test tailors the item difficulties to the individual child's ability level regardless of whether the child is functioning at grade or age level. In a heterogeneous national sample many children are far above or below typical grade-level performance yet they must be measured with relatively equal precision. Strict adherence to grade appropriate item pools has little relevance to measuring growth in longitudinal studies.

How Well Did the Adaptive Testing and the Multiple Criterion Referencing Work?

In addition to yielding relatively high reliability estimates, adaptive tests should help to reduce the typical finding of a relatively high and sometimes spurious negative correlation between initial status and gains. That is, that portion of the negative relationship that is due to ceiling effects should be reduced by adaptive testing. Adaptive tests should not and could not reduce a negative correlation between initial status and gains where that correlation has a structural basis. For example if a classroom teacher is only targeting the slower children in her class we would expect a negative correlation between initial status and gain regardless of whether the adaptive test is successful at minimizing floor and ceiling effects. Also the items marking the upper end of the test score scale appear to be more difficult in the sense that they are more language based and reflect the transition from basic phonics skills to a higher order of thinking and comprehension skills. These latter skills are probably more difficult to teach than the earlier based pre-reading skills and one would not expect children who make their gains in the upper levels of the scale to have gains as large in absolute value as those children making gains in the middle or the lower end of the scale.

Table 1 below presents the relationship between initial status on the reading test and gain for (a) the kindergarten year, (b) spring kindergarten and fall first grade, (c) fall first grade and spring first grade, and (d) spring first grade and spring third grade.

Inspection of Table 1 suggests that in the early periods of schooling (i.e., kindergarten through first grade) children are making relatively equivalent gains regardless of their initial position on the vertical scale. However, this is not true of the gains taking place between spring first grade and spring third grade. The negative correlation suggests that, over the greater time

Table 1

Correlations between Initial Status and Gains in Reading for Four Succeeding Time Periods

Fall-K status with gains between Fall-K and spring-K	Spring-K status with gains between spring-K and fall first grade	Fall first grade status with gains between fall first grade and spring first grade	Spring first grade status with gains between spring first grade and spring third grade
r = .16	r = .20	r = .07	r = –.39
N = 15286	N = 16041	N = 16264	N = 16295

span children at the lower end of the vertical scale are more likely to be making larger gains in terms of total scale score points than those children at the higher end of the scale. The question becomes is there a psychometrically imposed ceiling effect in the adaptive test partly because of the longer gap in time being spanned and/or is the third grade material measured at the high end of the scale inherently more difficult leading to an apparent slowing down of many children's cognitive performance with respect to the more complex material? It is after all during this schooling period that the reading material makes a transition from predominately phonemic awareness to reading comprehension. Some additional light can be cast on this phenomenon by looking at where the gains are taking place with respect to the hierarchically ordered criterion referenced scale points and the ascending skills reflected in these behavioral anchors. There are eight ascending criterion referenced points. Table 2 shows the various skills marking the ascending criterion referenced points along the vertical score scale and the percent of children making their gains in the area of each of the skill levels. This classification of gains is referred to as locus of maximum gains. The locus of maximum gains (LMG) is defined as:

$$\text{Max}\ (P_{i2} - P_{i1}) \quad i=1,\ldots,8 \qquad (1)$$

where: P_{i2} is the probability that given a child's latent trait score, theta, at Time 2, the individual will get three out of four items correct in the i^{th} cluster of items marking the i^{th} criterion referenced point. P_{i1} refers to the corresponding probability based on the child's theta at Time 1. If there are eight ascending in difficulty criterion referenced points and each of which is marked by a cluster of four items, each child will generate eight differences in probabilities depending on

how much the given child's latent trait score, theta, changes from Time 1 to Time 2. The LMG is that criterion referenced point that shows the maximum difference in probabilities.

Table 2 below contrasts the time period from spring K to spring first grade with the spring first to spring third grade with respect to the percentage of children making their maximum gains at each of the eight criterion referenced points on the scale.

Table 2

Percentages of Children Making Their Maximum Gains at Each of the Criterion-Referenced Score Points During the Periods Spanning Spring K to Spring First and Spring First to Spring Third

Skill levels	Percentage Spring K—Spring 1	Percentage Spring 1—Spring 3
Letter recognition	5.3	.9
Beginning sounds	8.2	.3
Ending sounds	16.8	1.1
Sight words	44.9	7.1
Comprehension of words in context	19.0	28.8
Literal inference	5.0	41.0
Extrapolation	.8	17.6
Evaluation	–	3.3
Totals	100.0	100.0

Inspection of Table 2 indicates that during the spring K to spring first grade year the vast majority of the children were predominantly making their gains in the phonemic awareness skills—letter recognition through sight words, while approximately 25% were gaining in beginning literacy skills comprehension of words in context through evaluation. In the case of gains between spring first grade and spring third grade almost 70% of the children were making their gains in lower level literacy skills (i.e., words in context and literal inference) while only 20.9% are making their maximum gains in the higher literacy skills of extrapolation and evaluation. These differences in percentages reflect for the most part what is emphasized in the curriculum that spans the first through third grades. There does not seem to be much empirical evidence for a ceiling effect since the children making their gains in the highest proficiency level

(evaluation) had a mean of 139.2 and a standard deviation of 3.7, which puts them about 2.3 standard deviations below a perfect score. Another way of looking at ceiling effects for the highest proficiency level is in terms of the skewness of their distribution of gain scores. The skewness index for the gain scores was –.07 with a standard error of .10. Ratios of skewness indicies to their standard error that are less than 2.0 are not considered significant (D'Agostino et al., 1990). It would seem that both the span of time covered, curriculum emphasis, and the complexity of the conceptual literacy skills required for gains at the very upper end of the scale contributed to the negative correlation between status and gain.

A Comparison of Traditional Adjusted Gains With the Difficulty-Coded Locus of Maximum Gain as Outcomes

Table 3 presents the regression of adjusted gains as well as the difficulty-coded locus of maximum gains on selected process and background variables for spring K to spring first grade. Table 4 presents the parallel regressions for the time period from spring first grade to spring third grade. The locus of maximum gains is coded 1 through 8, reflecting their difficulty hierarchy. That is, children making their gains on the vertical scale anchored by letter recognition marker items are coded 1 while children making their gains on the vertical scale in the area of evaluation marker items are coded 8, and so on.

Inspection of the standardized regression weights (betas) in Table 3 for the spring K to spring first grade year comparison indicates that the typical process variables such as highest working parent's education level (of the working parent), how often read to the child, whether attending a public or private school, highest degree expected and whether child was in a center-based program have consistently higher relationships with where the gains are taking place on the vertical scale rather than with how much the child is gaining. This is particularly true for highest parent education, which turns out to be almost three times as important for explaining where the gains take place in contrast to explaining the amount of gain. Similarly how often parent reads to the child is more than twice as important for where the gain takes place compared to the amount of gain. Also a comparison of the R^2's indicate that the complexity of the skill level being learned is much better explained by the process variables than is the adjusted amount of change.

Table 3

A Comparison of Traditional Gain Score Analysis Regressions With a Procedure That Emphasizes Where the Gain Takes Place for the Spring K to Spring First School Year

Model	Regression adjusted gains			Regression with difficulty-coded locus of gain as outcome		
	Beta[a]	*t*	Prob.	Beta[a]	*t*	Prob.
Parent educ. level	.120	13.210	.000	.309	37.325	.000
Often read to child	.050	5.990	.000	.124	16.132	.000
Degree expected	.039	4.729	.000	.016	2.032	.042
Public vs. private[b]	.067	8.059	.000	.078	10.100	.000
Child ever in center-based care[c]	.011	1.377	.169	-.057	-7.589	.000
Initial status	.147	17.325	.000	---	---	---
	R^2 = .075		P = .000	R^2 = .169		P = .000

[a] Standardized regression weight. [b] Private school coded 1, and public coded 0. [c] Yes is coded 0, and no is coded 1.

Table 4 presents the parallel regressions for the gains spanning spring first to spring third grade.

A comparison of the standardized regression weights (betas) in Table 4 for the gains taking place between spring first and spring third grade year show similar but slightly less extreme differences favoring the locus of maximum gain outcome. Once again the comparison indicates that the typical process variables such as highest working parent's education level (of the working parent), how often read to the child, whether attending a public or private school or ever in center-based care have higher relationships with where the gains are taking place on the vertical scale rather than with how much the child is gaining. The ever in center-based care variable was not significantly related to the *amount of gain* for both time periods but had a small but significant relationship with the *difficulty of gain* outcome over both developmental periods.

Table 4

A Comparison of Traditional Gain Score Analysis Regressions With a Procedure That Emphasizes Where the Gain Takes Place for the Period From Spring First to Spring Third

Model	Regression adjusted gains			Regression with difficulty-coded locus of gain as outcome		
	Beta[a]	*t*	Prob.	Beta[a]	*t*	Prob.
Parent educ. level	.210	24.909	.000	.331	40.309	.000
Often read to child	.067	8.780	.000	.134	17.572	.000
Degree expected	.052	6.950	.000	.020	2.589	.010
Public vs. private[b]	.006	-.817	.414	.073	9.469	.000
Child ever in center-based care[c]	.012	1.651	.099	-.040	-5.286	.000
Initial status	-.485	-61.335	.000	---	---	---
	R^2 = .204		P = .000	R^2 = .182		P = .000

[a] Standardized regression weight. [b] Private school coded 1 and public coded 0. [c] Yes is coded 0 and no is coded 1.

It is interesting to note that attending a public rather than a private school has little or no relationship with the *amount of gain* during the first to the third grade year, but has a significant positive relationship with the *difficulty of the gains*. This finding results from the fact that a greater percentage of the private school children (36% vs. 18%) when compared to public school children are making their gains in the skills represented at the upper levels of the test score scale. These advanced comprehension proficiency levels marking the upper ends of the scale not only reflect greater cognitive demands but also reflect beyond grade level knowledge and to a certain extent what is not yet being typically taught during these school years. As pointed out earlier one should probably not expect the *amount of gain* in absolute score points in these more demanding skills to be as large as those taking place at lower levels on the test score scale. An analogy here

would be a bicycle race in hilly country where the leaders are climbing a hill while their pursuers are still in the flat land below. They (the pursuers) will gain on the leaders at this point in time, but when they get to the hill they may fall further behind. The assumption of equal units on the vertical test score scale simply does not hold for most tests whether they are IRT-based and scored or whether they are simply using observed number right scores. This measurement problem has little effect on assessment of a child's status on a particular occasion, but it does make the interpretation of the absolute size of gain scores problematic.

While the R^2 in Table 4 for the traditional gain regression is slightly higher than that for the locus of gain, the majority of the predictable variance is primarily due to the pretest, which plays the role of covariate in the adjusted gain scores. In fact the increment in R^2 for the process variables after controlling for the covariate is only .05.

It would seem that process variables are more likely to have stronger relationships with the cognitive demand level of the skill being learned rather than with the amount of gain. They may also demonstrate a different pattern of relationships with the locus of gain outcome when compared to the amount of gain outcome. This potential for both differences in strength as well as pattern of relationships for the two gain outcomes is likely to be accentuated when the mode of measurement is adaptive. Since adaptive tests allow gains throughout the ability/achievement distribution, process and background variables, which tend to be related to ability (e.g., parents' educational level) are less likely to demonstrate strong relationships with *amount of gain.*

Conclusions

It is argued here that the appropriate measurement of change requires an adaptive test to minimize floor and ceiling effects. In addition, to arrive at a complete picture of change one must not only measure how much change has occurred for each child but also where on the vertical scale the child is making gains. The need for this latter measure of change arises partly from the lack of equivalency of test score scale units. Also quantifying where on the score scale the change is occurring can be made more policy relevant if the vertical test score scale is behaviorally anchored at ascending points with items reflecting learning milestones of increasing complexity. Then the locus of each child's gain can be identified with a specific learning milestone. It is further shown that at least in the early school years background and process variables tend to have higher relationships with where on the score scale gain is taking place rather than with the amount of gain. The more adaptive the testing environment the more likely

that the locus of maximum gain outcome can provide process relevant information above and beyond that provided by traditional gain score analysis.

References

Cleary, T. A., Linn, R. L., & Rock, D. A. (1968). An exploratory study of programmed tests. *Educational and Psychological Measurement, 28,* 345–360.

D'Agostino, R. B., Balanger, A., & D'Agostino, R. B. Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician 44*(4), 316–321.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Goldstein, H. (1998). MLwin [Computer software]. Bristol, UK: University of Bristol, Centre for Multilevel Modelling.

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congden, R. (2001). HLM [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.