Individual Growth and School Success

April 2004

Martha S. McCall
G. Gage Kingsbury
Allan Olson

A technical report from the NWEA Growth Research Database

NWEA
Northwest Evaluation Association
Partnering to help all kids learn

Northwest Evaluation Assocuation
5885 SW Meadows Road, Suite 200
Lake Oswego, OR 97035-3526

www.nwea.org
Tel 503-624-1951
Fax 503-639-7873

Individual Growth and School Success

Martha S. McCall, G. Gage Kingsbury, and Allan Olson

In virtually every state and most school districts in the country, students are now asked to meet rigorous performance standards in various academic areas. These performance standards are designed to assure that our schools are producing students capable of competing successfully in a global marketplace. Students who pass the standards should have the ability to go on to higher education or into the work force with skills that allow them to excel in their chosen field. NCLB has hastened the trend that was already underway to establish rigorous, consistent standards. Unfortunately, the existence of curriculum standards and performance standards alone is not sufficient to enable us to identify successful students and schools.

The primary component of AYP is the percentage of students being judged as "proficient" or better on the required tests. To be identified as making adequate yearly progress, a school must have a certain percentage of students determined to be proficient in each identified subgroup of students. To allow the states the ability to customize the requirements, NCLB requires the states to:

- establish their own definitions of proficiency
- establish their own tests to determine whether students are proficient
- establish their own cut-off scores on their tests to identify proficient students

These state responsibilities allow for variability in what is meant by "proficient" from one state to the next while assuring consistent comparisons for schools within states (Kingsbury, Olson, Cronin, & Hauser, 2003). This may, in turn, result in substantial difference in what type of school is identified as meeting AYP. While this flexibility is written into the law, it may result in several interesting results having more to do with statistics than with learning. Several of these statistical issues are discussed below, with theoretical and practical examples.

AYP measures should provide a complete, accurate picture of the effectiveness in the schools being judged. Whenever we reduce the variety in a set of data, information is lost. When all possible test scores are reduced to two data points (meets or does not meet standard) information about each student is lost. As we go from individual test scores measured at one point in time, to individual proficiency estimates at one point in time, to school summary percentages, we lose information at each step. It is unclear whether the remaining information is complete and accurate enough to provide public accountability and guide parents in choosing schools.

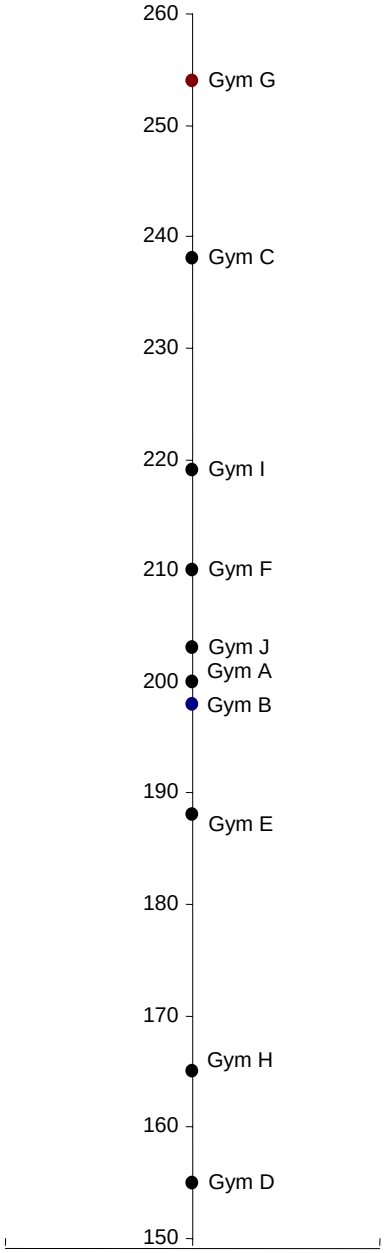
The current study investigates the effect of adding an individual growth measure to the primary definition of AYP. A variety of approaches to measuring growth have already been identified and are in use in schools across the country. In almost every statewide assessment system, a measure of the amount that individual students change from one year to the next is a feasible addition. The primary point of this study is to investigate whether, and to what extent, current AYP definitions could be improved by adding an individual growth measure to reflect the percentage of students who are currently identified as proficient.

This study considers how school effectiveness can be determined using student test results. To introduce some basic concepts, we begin with an extended analogy:

The Parable of the Gyms

In an attempt to promote physical fitness, a state governor has decided to rate gyms devoted to weightlifting. Consumers can use the ratings to decide which gym will be most effective in promoting fitness. Each gym has collected data on the amount of weight its customers can lift when they join the gym and how much they can lift after a year of training.

Figure 1.
Average Dead-weight Lift (in Pounds)



First, the governor adds up all the weight lifted at the end of the training year. Since some gyms have more members than others, he divides the weight lifted by the number of customers to get an average. He then sorts the list of gyms by average weight lifted.

Table 1.

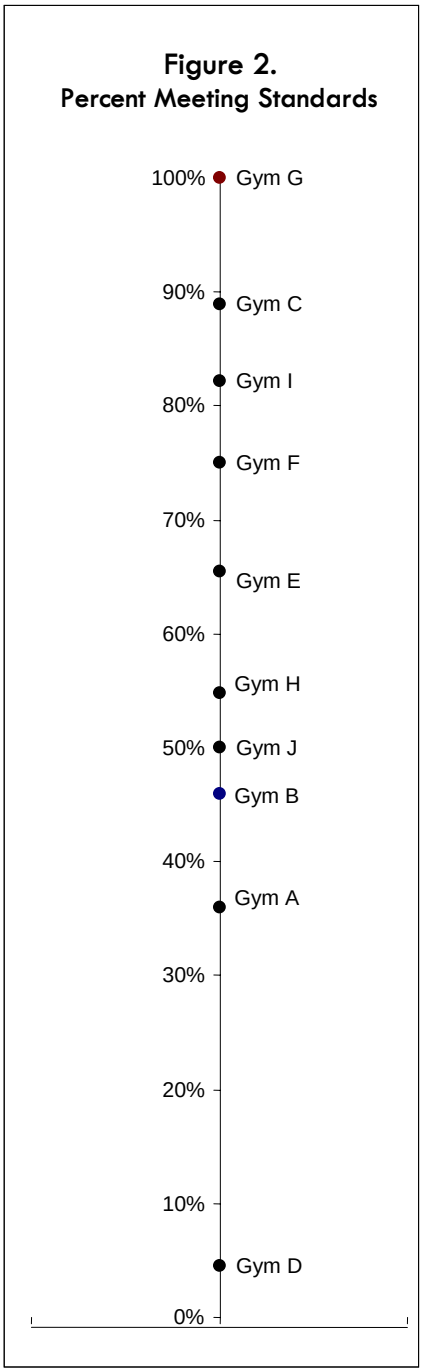
	Ave Dead-Wt Lift (in pounds)
Gym G	254
Gym C	238
Gym I	219
Gym F	210
Gym J	203
Gym A	200
Gym B	198
Gym E	188
Gym H	165
Gym D	155

The governor sees that Gyms C and G are clearly ahead of the others. Gyms J, A, and B form a close cluster, with Gyms H and D trailing behind. Should he simply draw a line at Gym F and say that all the gyms below it are ineffective? That seems too arbitrary.

He has his staff compute the average of all weight lifted. It turns out to be 203. If gyms with below average amounts of weight lifted are deemed ineffective, Gym J (203) would get a good rating, while Gym A, (200) just below it would not. Is there really that much difference between them?

The governor also ponders information about the lack of fitness in the overall population. Exceeding the average of a weak group is not a good goal. He wants to raise the fitness of the state's citizens, not accept their current status. Knowing which gyms are at or above average wouldn't tell prospective clients about whether or not the gyms actually succeed in making their customers fit.

The governor learns that a state health panel has published fitness guidelines based on the ratio of body weight to weight lifted. Using these, the governor asks his staff to calculate how many of each gym's members had met the criteria. (Conveniently, gym members weigh themselves when they check in.) He figures out the percentage of members meeting fitness standards for each gym. Again he arranges the results on a number line (See Figure 2).



He is sure that this is a fairer representation of the effectiveness of the gyms. He is now ready to look at the actual names of the gyms so that he can begin awarding ratings.

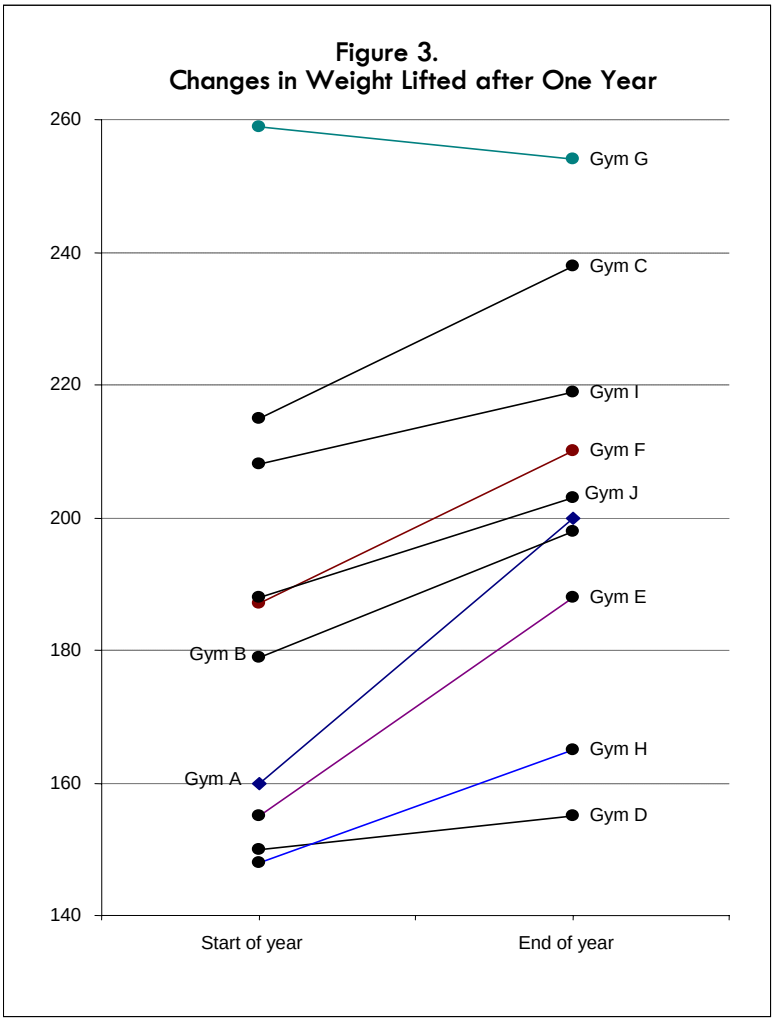
His trusty staff prints out the names of the gyms. After reading them, the governor reconsiders his system.

Code	Name
Gym A	Couch Potatoes Anonymous
Gym B	Big Jim's Big Gym
Gym C	Tough Guys, Inc.
Gym D	Mac's Free Weights and Donut Shop
Gym E	Steel Magnolias: Ladies' Fitness Experts
Gym F	G. Snooty's Elite Executive Spa
Gym G	Hod Carriers' Local #34 Rec Hall
Gym H	Oak St. Senior Center Strength Training
Gym I	Bob's Neighborhood Workout Place
Gym J	Pumping Irony: The Postmodernist Gym

He realizes that the results he has been examining may indicate the nature of a gym's patrons rather than the gym's effectiveness. The members of Gym G are likely to perform at a high level when they join the gym. Maybe the gym didn't do anything for them at all. It would mislead consumers to give it a high rating if this were the case.

The governor struggles with this problem. He wants all the citizens to be fit. He doesn't think that Gym F with the latest electronically-monitored equipment and a staff of personal trainers should have different standards from the others, but he knows that its clients come to it with a lifetime of healthy habits and the advantage of home equipment. He doesn't feel that he has enough information to know whether the gym, despite its relatively high ratings, is helping its clients to become more fit. He feels that consumers need more information before choosing a gym.

The governor uses his considerable charm to induce his staff to display increases in weight lifted in the year that patrons have used the gyms. This reveals some interesting patterns. Sure enough, the hod-carriers, whose jobs require them to lift loads of brick and mortar for long periods of time, had come in with the capacity to lift heavy weights.



Indeed, further investigation revealed that they completed the dead-weight lifts as a cool-down from work and used most of their gym time increasing flexibility under the tutelage of an attractive yoga instructor. Although its patrons reported greater feelings of well-being, Gym G should not be labeled as effective for the public.

Gym A is another interesting case. It has clearly been effective in raising the overall fitness level of its members and would seem to deserve a good rating. But Figure 2 shows that its patrons are still not fit. Should it be portrayed as a good model?

Table 2. Change in Weight Lifted in One Year

Code	Name	Difference
Gym A	Couch Potatoes Anonymous	40
Gym B	Big Jim's Big Gym	19
Gym C	Tough Guys, Inc.	23
Gym D	Mac's Free Weights and Donut Shop	5
Gym E	Steel Magnolias: Ladies' Fitness Experts	33
Gym F	G. Snooty's Elite Executive Spa	23
Gym G	Hod Carriers' Local #34 Rec Hall	-5
Gym H	Oak St. Senior Center Strength Training	17
Gym I	Bob's Neighborhood Workout Place	11
Gym J	Pumping Irony: The Postmodernist Gym	15

The governor makes a chart of the difference between weight lifted at the beginning and end of the year (Table 2). It shows Gyms A and E to be the most effective and Gym G the least. The governor is aware that it is difficult for people working at the limits of human ability, like the hod-carriers, to increase their performance whereas the reformed sofa jockeys at Gym A can more easily make dramatic changes from their low initial levels. Furthermore, the growth expected of seniors in Gym H may not be the same as that expected at other gyms. Clearly, the amount of change needs to be compared to the amount of gain expected for other people who began at about the same age and level of ability.

The governor now knows that he needs several pieces of information about each facility in order to inform the public: performance with respect to a standard, the pattern of growth over time, and a comparison of growth to a standard that takes the starting level into account.

We will leave the governor to ponder these factors while we draw parallels with educational evaluation.

Educational Applications

Like our fictional governor, educational evaluation has passed through stages of understanding about the best way to measure effectiveness. Until the last decade or so, most evaluation expressed results in terms of distance from average performance. That is how most people with assessment training were taught to portray performance and it was the only way most citizens could make sense of them. However, psychologists have known for quite some time how to construct cognitive scales that are independent of averages. Most people are familiar with these measures because of large-scale tests like the SAT, which gives scores that have the same meaning year after year even as the population moves up or down on the scale.

Mental skills like reading and mathematics cannot be observed directly or measured with a physical instrument like a ruler. Instead, the existence of an ordered set of skills is inferred from behavior. The branch of psychology and statistics that deals with assessment provides statistical tools for constructing stable scales and measuring where people fall on these scales (Embretson & Hershberger, 1999; van der Linden & Hambledon, 1997). Tests based on these methods have become far more prevalent because of the advent of high speed computers, advancements in statistical methods and a public demand for different types of information.

The educational reform movement was based on a belief that averages were too low and that there was a need to articulate the content and rigor of educational goals for all students. Furthermore, educational institutions had a societal responsibility to show the public how well students were learning in terms of goal attainment, not just in comparison to national norms. (Commission on the Skills of the American Workforce, 1990) The need to specify a level of attainment in terms of a minimum test score led to the development of standards-setting processes that assumed the existence of an underlying scale reflecting both the difficulty of test questions and the level of student achievement. (Cizek, 2001)

These ideas are prevalent in educational and policy circles throughout the United States. The federal Educational and Secondary Education Act (ESEA) has been a reflection, not a cause, of the change from norms-based to standards-based assessment. Many states had already implemented educational reform laws or policy prior to the 1994 Improving America's Schools Act (IASA) and well before the 2001 authorization of ESEA, the No Child Left Behind Act (NCLB). The US Department of Education has a history of responsiveness to trends and advances in public analysis and reporting. Now that state data collection systems have become more complete and assessment information is to be collected in adjacent grades (3 through 8), the data necessary for longitudinal comparisons is available. The nation is now at a point where policymakers and educators understand the need for growth-based measures and have the technical means for providing them.

Data Used in This Study

Northwest Evaluation Association constructed the Growth Research Database from assessments administered across the nation from 1996 to the present. It maintains longitudinal information about students, tests, and items and provides a tool for investigation of change over time. Records were extracted for this study if the academic subject was either reading or mathematics, the score was valid, the test was administered between April 10, 2002 and May 20, 2002 or between April 10, 2003 and May 20, 3003 and the grade level was between 2 and7 in 2002 or 3 and 8 in 2003. Students with valid 2003 scores were used in the study. Test records from 2003 were matched with 2002 records for the same student and subject.

Not all students with valid 2003 scores had matching records for 2002. In the results below, cross-sectional numbers for Spring 2003 are based on all students who took a test in 2003. Raw growth and growth index information is based on the students with valid scores in both 2002 and 2003. Only schools with at least 30 students taking tests at two points in time were included in the study. All test scores came from the computer adaptive Measures of Academic Progress test.

Table 3. Distribution of Schools and Students by Grade

MATHEMATICS			
Grade in 2003	Students with Sp '03 Scores	Students with '02 and '03 Scores	Number of Schools
3	19,301	15,850	261
4	20,814	16,686	275
5	30,693	24,777	374
6	32,261	23,701	234
7	30,124	23,146	172
8	18,259	14,843	118
Total	151,452	119,003	1,434

READING			
Grade in 2003	Students with Sp '03 Scores	Students with '02 and '03 Scores	Number of Schools
3	19,797	16,103	263
4	20,143	16,186	260
5	30,730	24,041	366
6	32,176	22,740	226
7	31,292	23,126	169
8	19,582	15,573	120
Total	153,720	117,769	1,404

The table above displays the sample break down by subject and grade. Records from 22 states were included in the sample. The distribution by state is displayed in Table 4. Note that the sample reflects NWEA's client base and is not intended to replicate U.S. demographics.

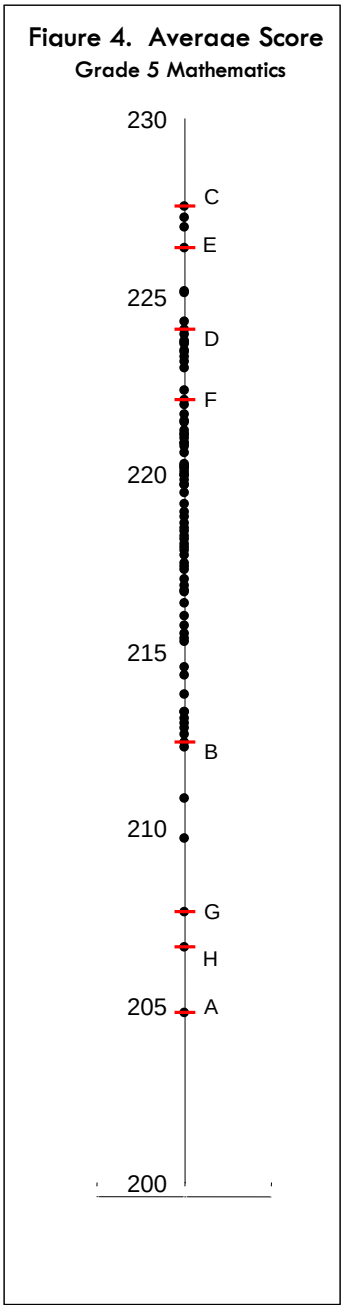
Table 4. Distribution of Sample by State and Subject

Mathematics			
State	Students with Sp '03 Scores	Students with '02 and '03 Scores	Number of Schools
AL	196	180	1
AR	1,308	995	5
AZ	366	277	1
CA	2,467	1,627	15
CO	9,211	6,778	47
IA	2,831	2,329	16
ID	25,010	19,537	133
IL	1,647	1,261	7
IN	71,296	57,263	273
KS	945	844	3
MI	3,004	2,323	17
MN	9,382	7,634	34
MT	2,815	2,402	13
NE	1,648	1,360	10
NM	3,463	2,507	17
OH	841	684	5
OR	164	103	2
PA	1,568	1,199	7
TN	5,541	3,643	16
WA	6,131	4,767	29
WI	1,499	1,183	11
WY	119	107	2
TOTAL	151,452	119,003	664

Reading			
State	Students with Sp '03 Scores	Students with '02 and '03 Scores	Number of Schools
AL	1,060	803	3
AR	1,326	1,004	6
AZ	845	608	2
CA	3,523	1,786	16
CO	12,348	9,288	60
IA	2,591	2,186	15
ID	25,559	20,216	135
IL	912	374	4
IN	67,587	53,494	254
KS	947	843	3
MI	3,295	2,528	18
MN	10,153	7,379	38
MT	2,754	2,209	13
NE	1,589	1,323	8
NM	3,226	1,930	16
OH	804	681	5
OR	245	197	2
PA	1,470	1,073	8
TN	5,588	3,686	17
WA	6,670	5,193	26
WI	1,146	893	7
WY	82	75	1
TOTAL	153,720	117,769	656

Some Examples

Figure 4 shows mean spring 2003 mathematics scores for a set of schools from a single state, Indiana. It is analogous to Figure 1, the average weight lifted in the gyms. The schools are ordered by rank, but it is not possible to know which schools are actually effective instructionally because we don't know the score that represents fifth grade math proficiency. Eight schools are labeled by letter in each graph to show how schools at different levels relate to the total data set when the data is displayed in a variety of ways. Of the eight, four will be examined in detail.



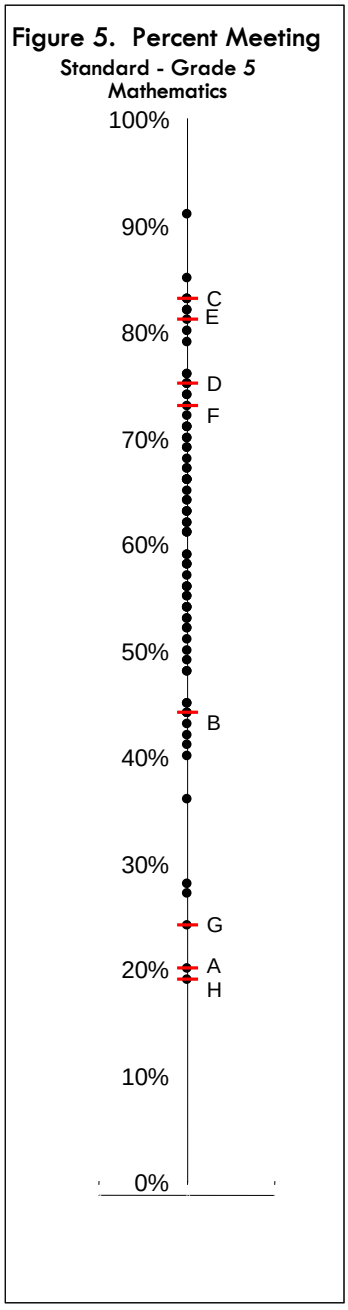
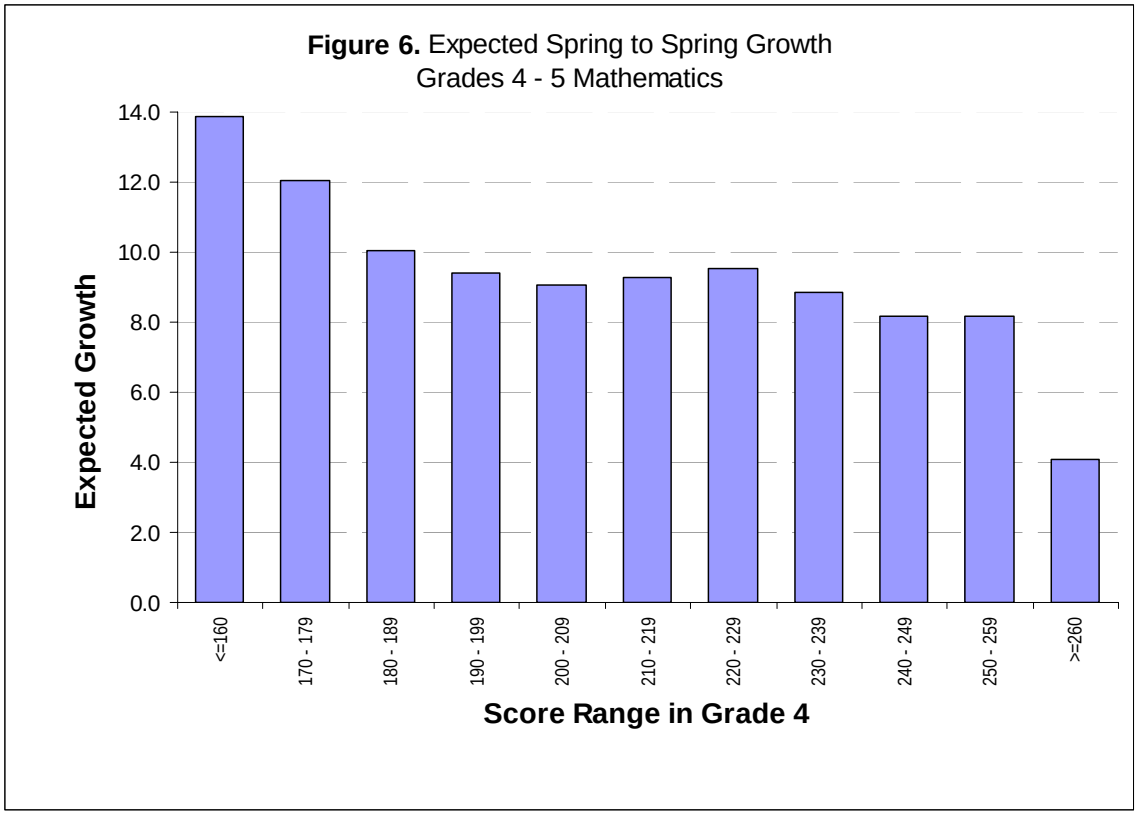


Figure 5 (analogous to Figure 2 above) gives more instructionally related information by displaying the percentage of students meeting state standards.

The numbers displayed here are results of selected schools on NWEA Mathematics achievement tests, not on actual state assessments. In a previous study (Kingsbury et. al. 2003), NWEA calculated state standard cut points on its reading and mathematics scales. These data are being used to illustrate general ideas about growth because they allow comparison across states and because the ideas can be seen more easily using a vertical scale. Indiana's cut score for 5th grade mathematics is equivalent to a score of 216 on NWEA's scale. This is a moderately high proficiency criterion falling at the 48th percentile of NWEA's spring norming population (NWEA, 2002).

Growth Measures

In the sports example introduced earlier in this document, the governor had the dilemma of deciding the amount of growth appropriate for those who started out as high performers compared to those who began at a lower level. He needed information about the normal growth of people at different initial levels; he needed growth norms. NWEA has conducted research to establish growth norms for its mathematics, reading, and language usage tests. There are sets of norms for expected growth at each grade from fall to spring, fall-to-fall, and spring-to-spring. The example here uses spring-to-spring comparisons (NWEA, 2002, p. 24) because this is the most likely comparison when using data sets for the No Child Left Behind Act. Figure 6 shows how much growth students usually make between spring of grade 4 and spring of grade 5. Note that students who begin at higher score ranges do not grow as much as those with lower initial scores.



NWEA subject matter scales are vertically scaled so that achievement at different grades can be compared on the same scale, much like a tape measure that records growth in height over time. Each school in our sample has at least 30 fifth grade students with mathematics scores in spring of 2002 and spring 2003. Raw growth for each student is computed by subtracting the 2002 score from the 2003 score. This is the absolute amount of growth over the year. Subtracting the expected growth from the raw growth shows how students did relative to others that began at the same level. This number is the growth index. Figures 7 and 8 display the mean raw growth and growth index for the Indiana schools.

Raw score growth can be computed for any sets of scores for different points in time. Because it is conditioned on grade and initial proficiency level, the growth index adds information to raw gains alone.

A number of models for representing growth are available. The analysis displayed here is an easily comprehensible method that exploits the vertical scale and growth norms for this particular data set. Other methods include longitudinal analyses of the percent of students meeting standard, changes in mean z-score or NCE (used in previous ESEA authorizations), regression of pre-score to post score and multilevel or hierarchical linear modeling (HLM). Flicek & Wong (2003) discuss these models in detail, noting their strengths and weaknesses. It is the purpose of this paper to illustrate the need for growth information in school accountability, not to recommend a particular growth model. The issues involved in choice of growth model center around precision, fairness, and comprehensibility. All growth models are based on the principle that tracking students over time while comparing rates of change to expected rates of change, yields information about how the school has contributed to student performance.

Figure 7. Raw Score Growth for Grade 5 Mathematics

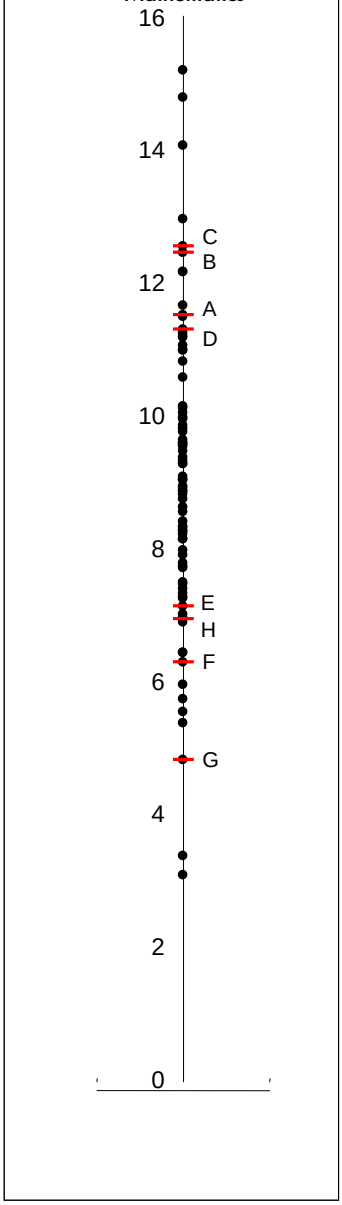
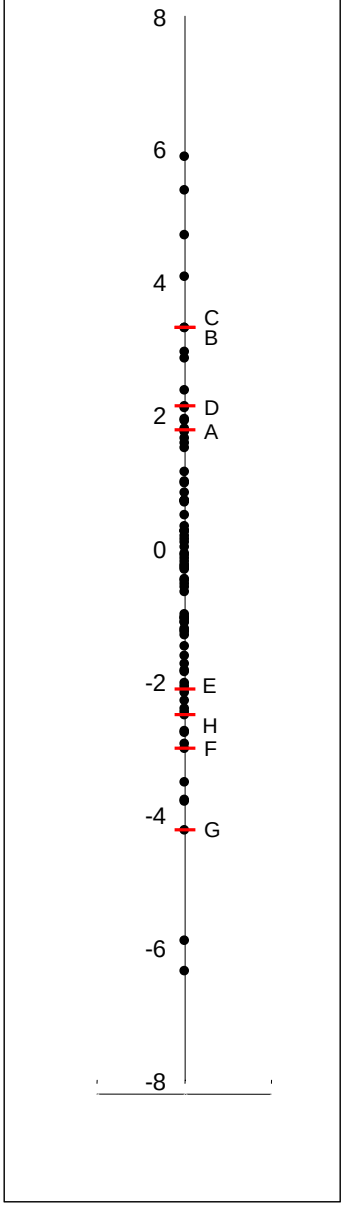


Figure 8. Growth Index for Grade 5 Mathematics



We now have two kinds of information about our sample of Indiana schools: information based on a single point in time (average score and percent meeting standard); information about change over time (raw growth and growth index). Displays of each of these are informative, but they are primarily useful for ranking schools. Combining single point and change data can give a richer picture of school performance.

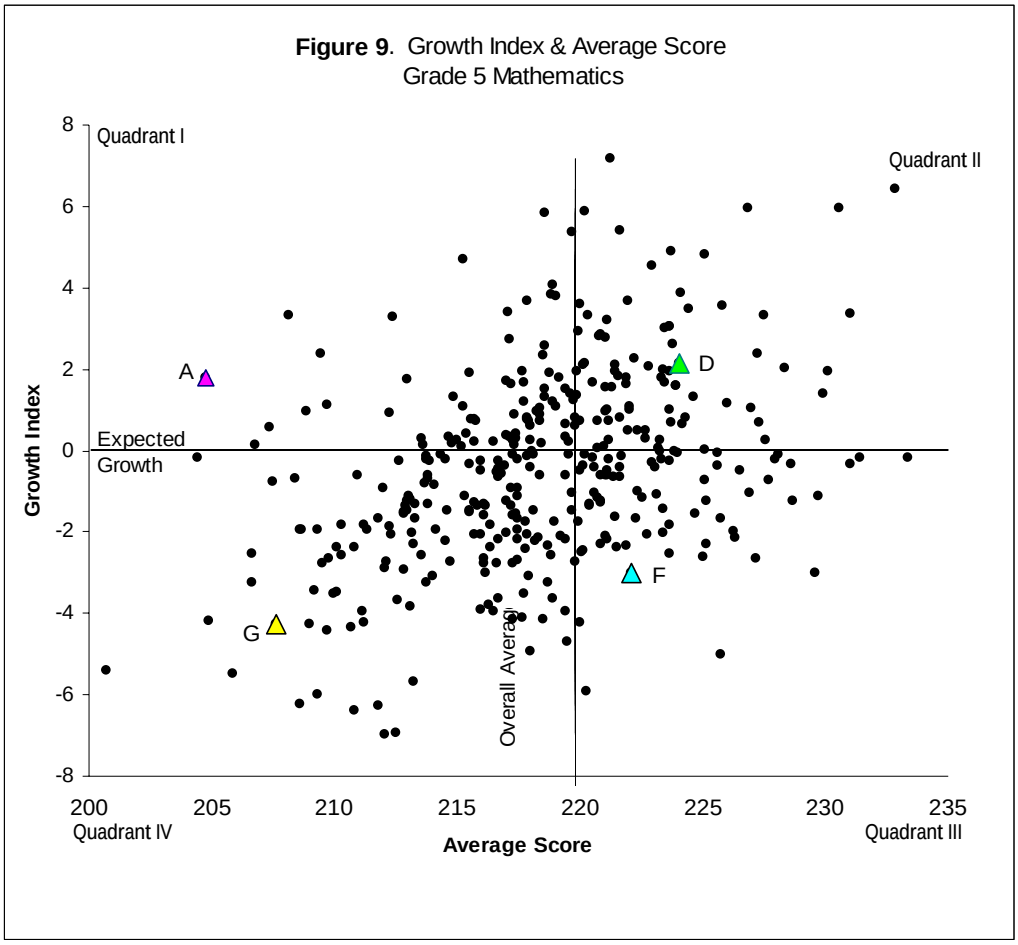


Figure 9 shows all of the study schools containing a 5th grade with mean spring 2003 score plotted against the growth index. The horizontal line represents expected growth; the vertical line is at the 5th grade mathematics norm (216.4). The quadrants contain samples representing high growth, low performance (Quadrant I), high growth high performance (Quadrant II), low growth, high performance (Quadrant III) and low growth, low performance (Quadrant IV). This example shows results for a single state, grade and subject. Appendix A shows the distribution of all schools in the study for each grade and subject.

Next, the characteristics of schools selected from each quadrant will be examined. The growth component adds information to cross-sectional data, i.e., results collected at one point in time. Figure 9 shows that schools A and B are in Quadrant I which contains schools with low test scores, but high growth. These are the Steel Magnolia Gyms of education, schools that have been effective at raising the skills of students with low achievement. Despite their exemplary growth, these schools are likely to fall below adequate yearly progress (AYP) requirements, eventually becoming subject to program improvement sanctions. Under program improvement, students from these schools are eligible to transfer to other, higher achieving schools.

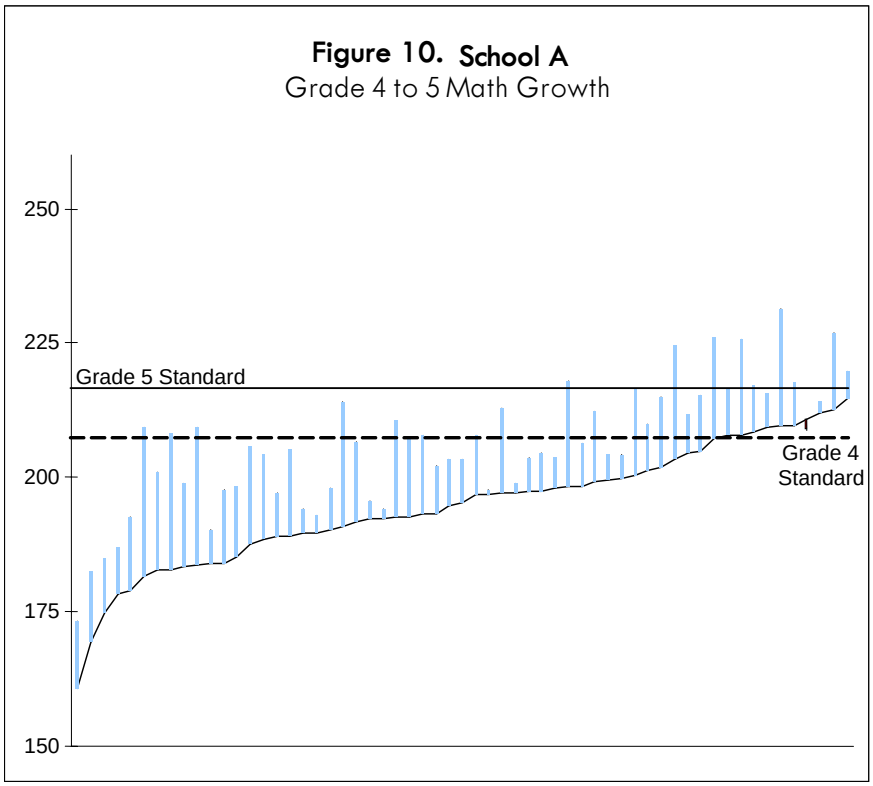
Schools in Quadrant II, including Schools C and D, have students with high performance and high growth. These schools take students with strong academic skills and challenge them to even higher levels.

Quadrant III schools also serve students with high achievement, but do not add greatly to the skill level. They are of particular interest in regard to the No Child Left behind Act because these schools often meet adequate yearly progress provisions and are thus eligible to receive students from schools that do not meet AYP. Would students in schools A and B be better off in schools F and E where they are likely to languish?

Finally, Quadrant IV schools are struggling with both low performance and low growth. These are the schools that need assistance leading to a coherent curriculum, better governance and instructional leadership. Shouldn't a meaningful evaluation model distinguish between schools in Quadrant I and Quadrant IV?

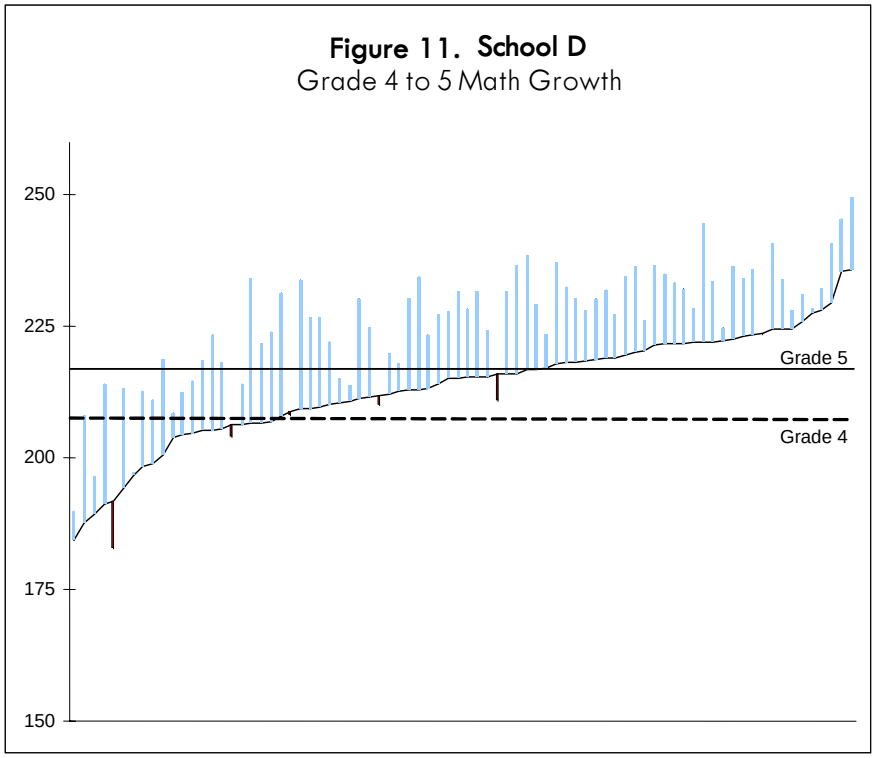
Tracking Individual Growth

We have identified schools of interest from each quadrant. The next section examines growth patterns for each quadrant in detail.



Figures 10 and 11 contrast School A, a low performing high growth school from Quadrant A with School D, a high performing, high growth school. The scale at the left is the mathematics scale. NWEA uses a vertical scale, one that extends across grades and allows the display of raw growth across time. Each vertical bar represents a student. The climbing solid line indicates the starting point in grade 4. Each bar begins at the solid line and ends at the 5th grade score. Horizontal lines indicate the NWEA derived equivalent standards for each grade. This gives us a visual referent for student performance. In both schools almost all of the student bars go up, representing a positive change. The lengths

of the lines are distributed fairly evenly across the continuum. That is, students of all levels seem to be growing academically, not just high or low performers.



Most of the students in School A however, are functioning below standard. There are some students who did not meet the grade 4 standard and grew enough to meet standard in grade 5. But most students are still below the bar and would not contribute to the school's success under NCLB. In School D, most students meet standard at both points in time in addition to exhibiting strong growth. At the end of grade 4, nearly half of the students were already performing above the grade 5 standard.

Now look at schools F and G, Figures 12 and 13 respectively. The growth bars are shorter in these schools. In school G, low performing students appear to have made the most growth, but those at higher levels have grown little or even lost ground. This school may have devoted a lot of resources to low achievers, but neglected other students. School F has low growth or academic loss across academic levels. However, unlike school G, most of its students have met both the 4th and 5th grade standards. Under the NCLB model, it would be a successful school. Figure 5 reveals that Schools D and F are nearly equivalent when you look at the percent passing standards at a single point in time. Yet they clearly differ in effectiveness.

Figure 12. School F
Grade 4 to 5 Math Growth

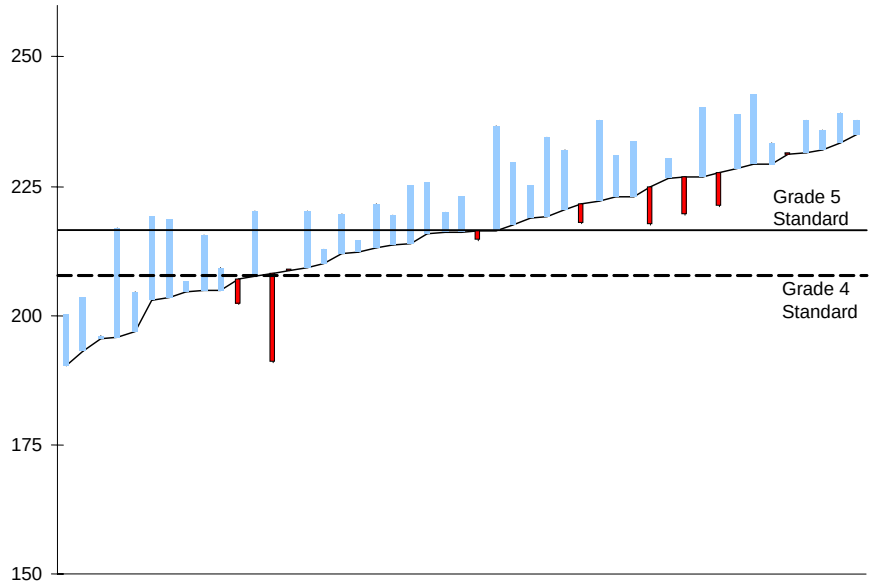
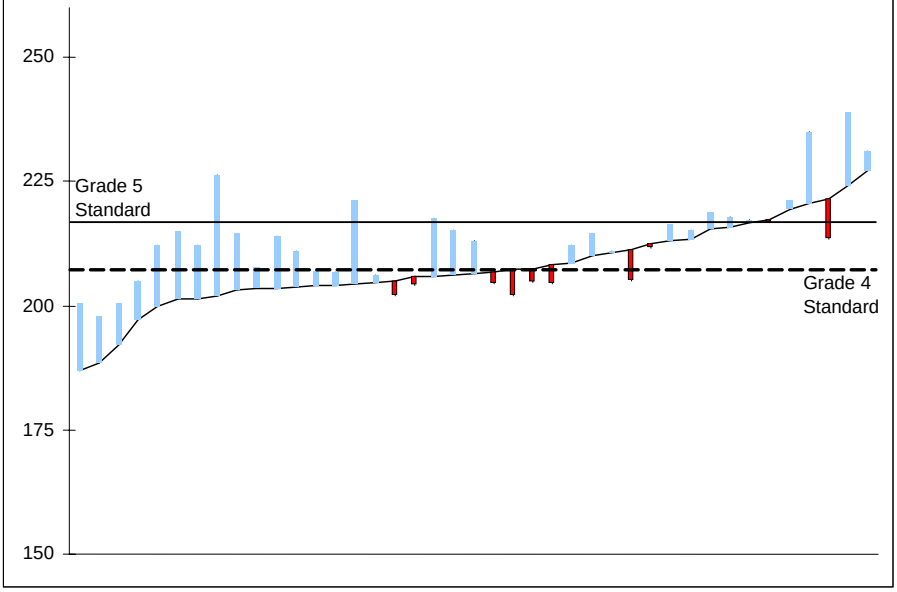


Figure 13. School G
Grade 4 to 5 Math Growth

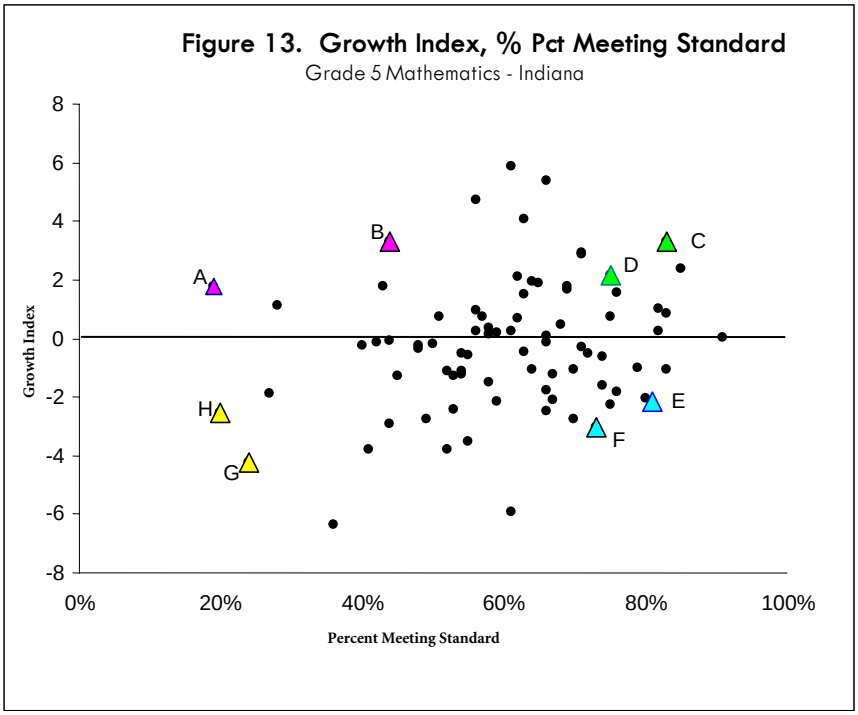


Finally, the charts below show observed and expected growth for individuals from each school. These are the same students displayed in Figures 10 – 13. This time they are arranged in order of their raw growth, represented by the vertical bars. The short horizontal line across or above each bar represents expected growth for the student. Expected growth varies depending on initial score (See Figure 6.), so there is some variation in the height of the horizontal bars. Figure 10 shows that school A had some

very low performing students in spring 2002. These students had extremely high expected growth compared to students who began at higher levels, so their expected growth bars float above those of most of the students. These graphs show how students actually grew compared to how we would expect them to grow.

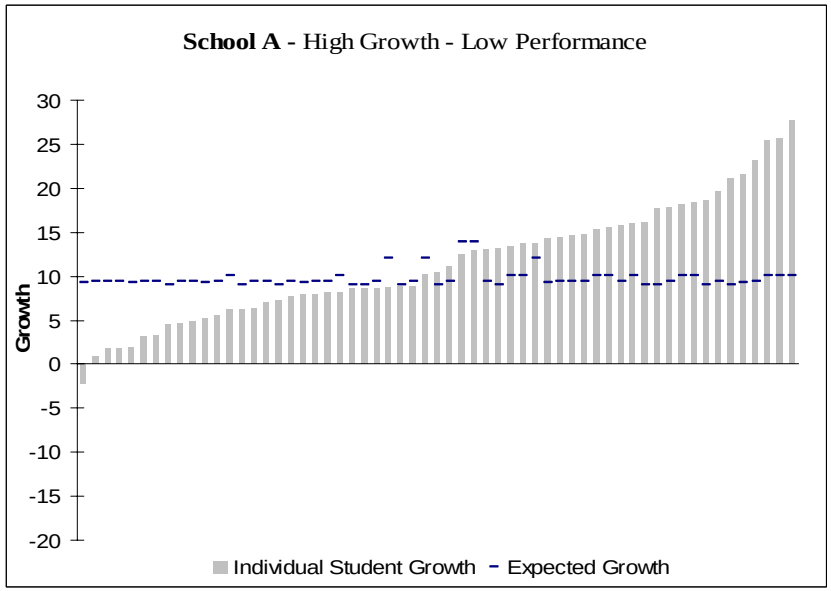
They show that students at schools A and D exceed expected growth more often than those in schools F and G. When you look across the chart, you see more vertical grey bars reaching or exceeding the expected growth marker. That is, growth in schools A and D is high when initial conditions and age (grade) are taken into account. This is what puts them into Quadrants I and II while schools F and G with fewer students meeting growth expectations are in Quadrants III and IV.

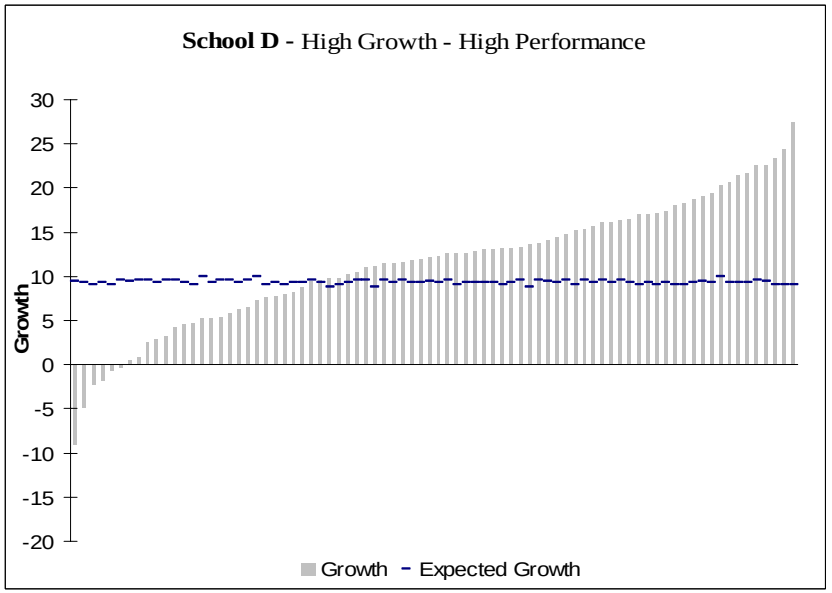
The equivalent cut score for meeting standard from Kingsbury et.al. (2003) was used to compute the percentage of students meeting standard in Indiana. As expected, when displayed with the growth index, the Indian distribution of our focus schools is very similar to that in figure 9 with some high performing schools showing low growth and vice versa.

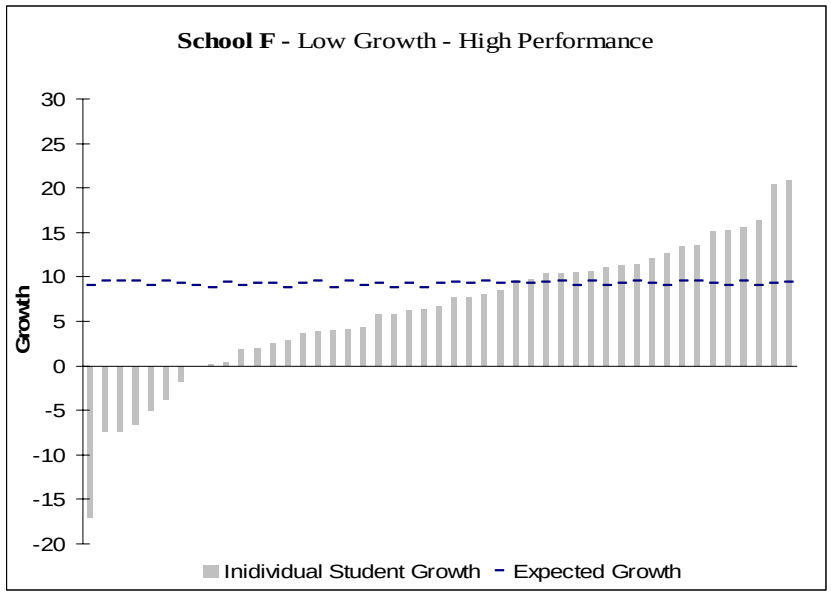


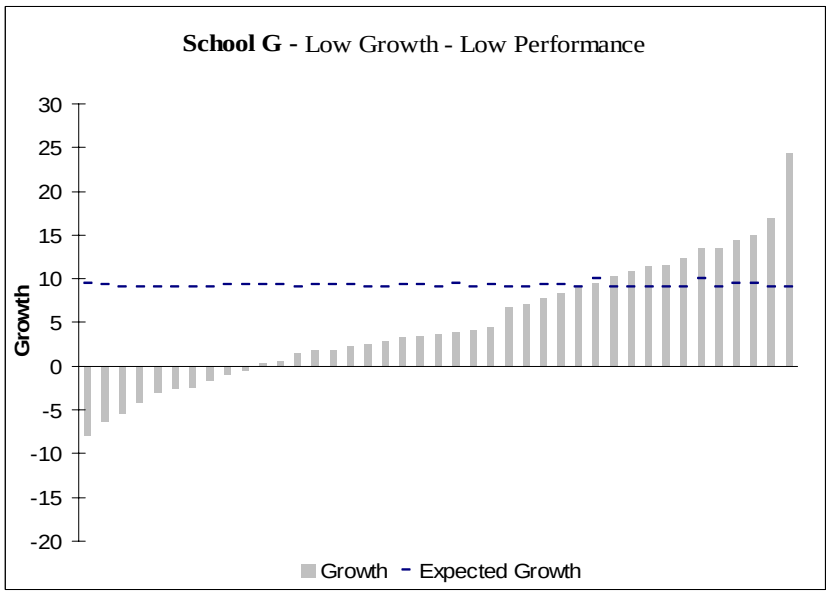
Again, we are able to show this graphically using NWEA's growth norms and vertical scale, but growth can be modeled by a variety of methods. Other models use different kinds of scales or compute growth expectations in other ways. However, the call for the use of growth as a component of school evaluation is consistent among educators, policy makers and measurement experts.

Figure 14. Expected Growth and Observed Individual Growth









Accountability and ESEA

It is the purpose of this paper to suggest that adding a student growth component will lead to a stronger AYP model, one that more accurately distinguishes effective from ineffective schools. NCLB regulations mandate a "snapshot" view of student results-cross-sectional percent meeting standard at a single point in time. Safe harbor provisions use snapshots at two points in time for different groups of students. Both methods are more of a measure of demographics than school effectiveness. Longitudinal or growth elements may be used, but only in a conjunctive sense. That is, growth may be used to identify more schools that have failed to make AYP, but not to mitigate cross-sectional results. This has not always been the case. Prior to NCLB, longitudinal models were the primary method for ESEA accountability. (See section entitled "The evolution of AYP" on page 24 of this document.)

Although the NCLB model addresses some previous criticisms, it has limitations. Previous longitudinal measures like NCE's and grade level equivalents were flawed because they failed to consider desired skills. The current measure is flawed because it fails to measure progress (or lack of it) over time. This paper suggests methods of analysis that preserve the high expectations of the educational reform movement, while providing more accurate information for differentiating between effective and ineffective schools.

LIMITATIONS OF CURRENT AYP IMPLEMENTATION:

1. Single point-in-time analyses may reflect demographics rather than effectiveness. They cannot distinguish between schools that accelerate skills and those that allow students to languish. Cross-sectional measures do not tell us whether students entered with high or low skills or whether they have gained or lost ground as a result of instruction. Flicek & Wong (2003) characterize the cross-sectional percent-proficient model as one of the least valid evaluation methods. Schools with high percentages of historically low-performing groups--students in poverty, ethnic minorities, Limited English Proficient students and students in special education programs—tend to be identified as failing to meet AYP more often than those with lower percentages of these groups. The performance of these students is reflected in the overall group performance and in each selected subgroup. Schools that serve primarily white, English speaking students who are not in poverty have higher results overall and frequently have subgroup numbers too low to report. The data do not show which schools have been effective with the population that they serve. (Kim & Sunderman, 2004b; Baker & Linn, 2002; Buchanan, 2004). Cross-sectional gains models, like the current safe harbor provisions, often end up measuring social differences in two successive groups of students rather than the effect of the school.

The table below shows the correlation between percent free and reduced lunch and cross-sectional mean score, mean raw growth and mean growth index for schools with available data. At every grade level and for both subjects, the correlation with the cross-sectional measure is negative (high poverty is associated with low performance) and significant. The correlation with growth measures is much weaker and frequently not significant. This means that schools with fewer economically disadvantaged students have higher scores overall, but they contribute to academic growth at about the same rate as schools with high percentages of students in poverty.

Table 6. Correlation Between Percent Free/Reduced Lunch and Achievement Indicators

READING				
Grade	No of Schools	Mean Spr '03 Score	Raw Growth	Gain Index
3	251	-0.608 **	-0.085	-0.270
4	244	-0.653 **	0.094 **	-0.136 **
5	338	-0.657 **	0.025 *	-0.186 **
6	207	-0.651 **	0.116	-0.064
7	149	-0.608 **	0.007	-0.127
8	104	-0.541 **	0.265	0.135

MATHEMATICS				
Grade	No of Schools	Mean Spr '03 Score	Raw Growth	Gain Index
3	249	-0.506 **	0.064	-0.019 **
4	259	-0.582 **	-0.166	-0.242 *
5	352	-0.589 **	-0.127	-0.143 **
6	214	-0.552 **	-0.054	-0.035
7	156	-0.527 **	-0.150	-0.125
8	104	-0.550 **	0.026 **	0.018

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Strong correlations between performance and socioeconomic indicators are often used to question test validity or to call for lower standards for students in poverty. We are not making either of those arguments. In this context, we are pointing out that schools with wealthier students may not be as effective as they appear in cross sectional analysis. It is not necessary to have different standards for different populations. By taking the initial score into account, each student acts as his or her own control group (Sanders & Horn, 1994; Thum, 2002; Hershberg et. al., 2004). The object of growth models is to find out how much the school contributes to results.

2. The NCLB model does not take the performance of students above or far below the standard into account. When the goal is to get the greatest number of students to meet the standard in a year, schools quite sensibly direct efforts at those performing just below the cut point. The model does not evaluate the progress of students who have already met standard. Schools and districts earn no credit for improving skills of the lowest performing students or for getting gifted student to work to their capacity. Indeed, critics have pointed to this feature of NCLB as a disincentive to excellence, encouraging states to set low standards in order to concentrate on fewer students and look better in public reports (Marion, et al, 2002; Hoff, 2002). The standard AYP measure looks only at two categories-those who meet standard and those who do not. It directs the focus away from rich score information about where students at every level are performing and what they need to do next. NCLB's safe harbor provisions look only at changes in the percent of students meeting standard, not at growth or stagnation throughout the school.

3. The current system does not necessarily lead to better placement for students in low performing schools. The examples shown above indicate that students who move to schools with higher percentages of students meeting standard may not get a better education. As Kim & Sunderman (2004a) note, students who take advantage of transfer opportunities afforded under NCLB often move from schools with support for low performing students to more affluent schools that do not have remedial reading programs, tutors or supplemental Title I money. Karl Meiner, a high school teacher in Portland, Oregon has taught at both inner city and suburban schools sees the results of these transfers. "The AYP system is punitive to schools that have disadvantaged students. NCLB erodes neighborhood unity by encouraging kids to travel across town rather than reviving and rehabilitating neighborhood schools."

4. Expectations of AYP need to be tempered by looking at observed results in exemplary schools. Lee's (2004) analysis of state data using current AYP provisions shows how unreachable the goals appear to be even when rolling averages and safe harbor provisions are used Robert Linn (2003), in his 2003 address as president of the American Educational Research Association, illustrated the gulf between NCLB expectations and observed performance. Using state and NAEP data from across the country, Linn projected that reaching 100 percent proficiency in twelve years would be highly unlikely. He called for the use of research to establish goals that are stringent, but feasible.

> "Objectives mandated by the accountability system should be ambitious, but also should be realistically obtainable with sufficient effort. It is not that current levels of student performance or that gains in student performance that typically have been achieved in the past are fine and should be adopted as the standard to be expected in the future. Rather, current levels of performance and past gains provide a context for judging future gains and long-range targets of performance."

He urges an examination of the highest performing schools and districts to find viable goals. These schools also form a set of exemplars to look toward for practices that lead to success.

PROPOSED MODEL AND HOW IT ADDRESSES AYP LIMITATIONS:

The intent of NCLB is to provide options for students in low performing schools. As more and more schools fail to reach the expected percent meeting standard, options for families become more limited. For most students there are few, if any, schools to transfer to and little incentive for their neighborhood school to improve when standards are unattainable (Kim & Sunderman, 2004a, 2004b). How can we find a system that does a better job of identifying effective schools, setting high but reachable goals and giving information about schools to emulate? We propose including growth-based evaluation in the AYP model.

Let us return to our fictional befuddled governor. He wants to know that gyms lead their patrons to fitness. In order to do this he needs to know what fitness standards are, how gym members have progressed across time, and what rates of progress are people can actually achieve given their age. Finally he wants to know what kind of success gyms have

with all of their clients, at any stage of strength. These are the elements of a good evaluation system:

1. A set of worthy goals
2. Use of empirical data as a guide to the possible
3. Everyone's performance contributes to judgment of the whole
4. Credit is awarded for reaching goals
5. A set of data tracking progress over time
6. A mechanism for reporting to the public

How can growth information be combined with the current cross-sectional model?

How do we achieve these in this setting? Almost every state has specified its set of worthy goals. Now that states have better data systems and are required (by 2006) to test in grades 3-8 and 10, the longitudinal data set should be feasible. How do we make the elements into a coherent system?

Complex mixed model methods (Thum, 2003; Sanders & Horn, 1994) treat all effects comprehensively. For individuals, the current score is affected by several past scores, membership in a school, nested within a district and a state. When the unit of accountability is the school, all of these factors for each child are used to compute an expected amount of growth. One of the advantages of these methods is the ability to partition growth into parts that are attributed to the school, district and to previous scores. They usually encompass several years of data for each analysis.

Currently, many states lack the years of data in all grade levels needed for NCLB analysis. When states have been testing at grades 3 through 8 for several years, more will be able to use these methods. There has been some controversy over whether a vertical scale is required for complex nested models (Hill, 2003; Bock, 1996; Thum, 2003). When a vertical scale is not present, data is scaled normatively. To be useful, results need to be tied back to content and scalar difficulty associated with standards-based cut points. There is broad agreement that the series of tests in the models must be based on the same construct.

Another area of concern is that the models are too complex to be understood thoroughly by policy makers and the public. (Indeed, Sanders model is a proprietary secret.) Simplicity is a consideration, but is not absolutely essential. People use a variety of complex indicators (the consumer price index, sports statistics, weather indicators) without knowing how to compute them. Thum (2003) recommends that models be open to expert review and replication, assuring the public that measures undergo professional scrutiny.

Compensatory index – Prior to NCLB many states combined snapshot data with growth indicators in a more direct manner than using the complex methods discussed above. (Marion, et al 2001; Seltzer, et al, 2002). Computation of these indicators is by no means easy (the Kentucky KIRIS method required a twenty-page manual) but it is straightforward and can be accomplished by anyone with patience and a four-function calculator. For many policymakers, this clarity helps dispel perception of secrecy on the part of public agencies.

The two-tiered approach – Hill (2003) proposes using a two-tiered model based on the safe harbor provision of NCLB. In the two-tiered model, a school that failed to meet the cross-sectional NCLB criteria, could use a growth approach as a safe-harbor model. If this provision were met, the school would make adequate progress. One of the appealing aspects of this model is that it could be used without changing the bill itself. Some rewording of the guidance would be needed, but no change to the law itself.

A proposal for a hybrid model – When a cross-grade (vertical) scale is in place, an intriguing model first offered by Kingsbury and Houser (1997), the Hybrid Success Model (HSM) can be pursued. In this model, a growth target is set for each student. The target is based on the student's current distance from a predefined proficiency level that is on the vertical scale. Both the student's absolute distance from the proficiency level and information provided by growth norms from similar students' performance on the same scale are used to moderate target setting. Students who are further away from the proficiency level will, of course, have higher growth targets than those students who are closer to the proficiency level. For students who are very far from the proficiency level, growth norms would be used to set a reasonable growth target that will place the student on track to meet a proficiency level two or three grades later. However, students who are very close to or already beyond the predefined proficiency level before instruction even begins will have targets set (moderated) based on the growth norms. The HSM growth targets will be challenging for all students, not just those who are below the proficient level. Moreover, because the HSM growth targets are conditioned on distance from the proficiency level, its use as an AYP model is not only consistent with the current intent but also allows consideration of a more complete range of student performance to be used in judging school adequacy.

Even if AYP doesn't change, schools and districts should give growth information in the spirit of consumer choice. As Kernan-Schloss (2004) notes, districts and schools can take charge of the public conversation by presenting more and better information, not by evading accountability. At the very least parents need to know what schools can offer and become voices for better education in every school.

Addendum

The Evolution of AYP

ESEA evaluation models have reflected social and measurement ideas current at the time of use and have changed to reflect public policy and statistical innovation. The central mission of ESEA has remained the same, but methods for carrying it out have evolved over time. Responsiveness to public input about Title I regulations has been vital to ESEA's long legislative popularity.

Early years of federal accountability. No Child Left Behind is the current authorization of the 1965 Elementary and Secondary Education Act (ESEA, P.L. 89-10). Enacted on the heels of the 1964 Civil Rights Act, ESEA was part of President Lyndon Johnson's War on Poverty. It was the first major source of federal funding for American education and was intended to assure that children in poverty could gain skills needed to thrive economically. ESEA began with a "compensatory" model. Federal funds were to be used to compensate for deficits caused by conditions of poverty in the home or in communities where schools were located (Schugurensky, 2002; Young, 2004).

In the 1960's federal aid to education was sponsored by a tenuous political coalition. Because of fears about federal control accompanying federal dollars, the survival of ESEA depended on making a minimal impact on state and district practices. Title I of ESEA was seen by Congress and the public as a way of equalizing funding between rich and poor schools and enforcing desegregation. At this stage, money was disseminated without a great deal of restriction on how it was spent and local programs were not evaluated as such. (Kaestle, 2001; Jennings, 2000).

By the 1970's, however, the public wanted federal dollars to go directly to needy children rather than being used for general system upgrades. In response to reports that funds were being used for general operating expenses, federal officials instituted detailed audit and compliance mechanisms to make sure money went to low performing students in low-income schools. During this time the focus was on providing access to services, not on evaluating their quality. Indeed, in the early years there was a prevailing belief that access would automatically lead to achievement (Natriello & McDill, 1999). This auditing system was in place during the 1970's and most of the 1980's as the major form of ESEA accountability. When President Ronald Reagan took office in 1980, he expressed his disappointment in ESEA, but did not change accountability at the school or district level.

Schools collected some data during this time, but it not used for public accountability. (This was accomplished through external contracted studies.) The social goal was to bring disadvantaged students to the same academic level as their non-disadvantaged peers. Therefore, the measurement model called for getting more students in low-income schools to perform at, or closer to, grade level. A program was thought to be doing a good job if its students were approaching the appropriate grade equivalent score. States and districts used standardized tests, usually from major test publishers with large norming groups. It quickly became apparent that grade-equivalent scores from different tests were not comparable. Programs were then required to convert scores to normal curve equivalents (NCE's) which are percentiles expressed on a different scale for ease in computation. Although relatively sophisticated for the time, given that programs of all

sizes and types had to apply the rules, NCE's didn't help comparability a great deal (Herman, Baker & Linn, 2001).

These evaluation designs were longitudinal in nature; they followed a cohort of students over time and compared initial and final stages. They assumed that expected growth meant a gain of one year in grade equivalence models or staying at the same NCE in normal curve equivalence models. Only students receiving Title I services were included in these evaluations. If results were better than expected, the program was credited with the increase.

ESEA begins to reflect reform ideas. Programs weren't required to set outcome targets for NCE gains until the Hawkins-Stafford Amendments of 1988. High poverty schools were permitted to become schoolwide programs, which based targets on the performance of all students in the school and enjoyed relaxed fiscal accountability. At that point, the education reform movement had begun in earnest. Policymakers knew that results were important, but were not as sophisticated about testing as they are now. Faced with the task of gathering data on a variety of instructional methods and settings, federal officials retained formulas based on traditional measurement ideas. Although the designs were normative in nature, educators became accustomed to looking for improvement in groups of students across time and were familiar with the logic of taking initial achievement levels into account when evaluating program effectiveness.

Critics pointed out that approaching, but not achieving, average performance preserved the achievement gap for students in poverty and that in any case there was no indication that average performance represented worthwhile skills (Commission on the Skills of the American Workforce, 1990; SCANS, 1991). Furthermore, analysis based on rank or distance from the mean is a zero-sum game. If one group gains in rank or position on a normal curve, another group loses. People wanted a design that treated knowledge as a resource that can expand without limit and be possessed by everyone.

To remedy these concerns, many states adopted policies requiring a definition of desired academic standards. States set standards representing the level of skills and knowledge students needed to thrive as citizens and workers. Performance standards did not necessarily reflect observed average performance. A guiding principle of the educational reform movement was that academic standards would be high and that all students would be expected to meet them.

The 1994 version of ESEA, the Improving America's Schools Act, mirrored the reform movement by requiring states having such policies to measure annual progress toward the goal of having all students meet academic standards. States used a variety of methods and degrees of rigor to define adequate yearly progress (AYP). Many states used progress over time as an AYP component. Because tests were required only at benchmark levels, and because most state data collection systems were relatively new, growth designs tended to be cross-sectional rather than longitudinal. These new growth measures were related to standards rather than to grade level, mean or percentile (Stecher & Arkes, 2001). This allowed states to express growth in relation to a scale or set of categories representing underlying skill criteria. In addition to making results more meaningful, the existence of criterion referenced achievement scales allowed academic progress to be measured in a way that could not occur in normative models where all results sum to zero.

The complexity and diversity of models and the lack of rigor in some states made policymakers impatient. The 2001 authorization of ESEA, the No Child Left Behind Act, sought to standardize AYP by defining it as a one-point-in-time, or cross-sectional, measure of the percent of students meeting state standards compared to a percent that would lead to 100 percent of students meeting standards in 2014. Schools, districts and states need to meet goals as a whole and for each selected subgroup specified in the law. States were no longer allowed to use their own, more complex, often more accurate models. Longitudinal growth models are not part of NCLB at all. Schools or districts that fall into safe harbor, which permits the use of cross-sectional gains in percentage of students meeting standard. It does not allow for longitudinal growth and does not measure the progress of students above and below the standard.

References

Baker, E. L. & Linn, R. L. (2002*). Validity issues for accountability systems.* CSE Technical Report 585, 2002. National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles: UCLA.

Bloom, H. S. (2002). *Measuring the impact of whole-school reforms: Methodological lessons from an evaluation of accelerated schools.* Planning and Evaluation Service, Office of the Undersecretary, DOC #2002-10. Washington D. C.: U. S. Department of Education.

Bock, R. D., Wolfe, R., Fisher, T. H. (1996). A review and analysis of the Tennessee value-added assessment system. Nashville: State of Tennessee.

Bucahanan, B. (2004). Defining 'adequate yearly progress'. *American School Board Journal*, February, 2004.

Cizek, G., Ed. (2001).*Setting performance standards: Concepts, methods, and perspectives.* Mahwah, N. J.: Erlbaum.

Commission on the Skills of the American Workforce. (1990). *America's Choice: High Skills or Low Wages.* Rochester, NY: National Center on Education and the Economy.

Drury, D. & Doran, H. (2003). The value of value-added analysis. *Policy Research Brief*, Vol 3, No. 1, 1-4. Alexandria, VA: National School Boards Association.

Embretson, S.& Hershberger, S., Eds., (1999). The new rules of measurement: What every psychologist and educator should know. Mahwah, N. J.: Erlbaum.

Flicek, M. & Wong, K. (2003). The challenge of using large-scale assessment to hold schools accountable. Submitted.

Herman, J. L., Baker, E. L., & Linn, R. L. (2001). Comparability: An elusive goal. *The CRESST Line.* Spring, 2001.

Hershberg, T., Simon, V.A. & Lea-Kruger, B. (2004). Measuring what matters. *American School Board Journal*, February, 2004.

Hill, R. (2003). *Using longitudinal designs with NCLB.* Dover, NH: Center for Assessment.

Hoff, D. J. (2002, October 9.) States revise the meaning of proficient. *Education Week*, 1, 24-25.

Holland, J. (2000). *Curiosities of Measurement in Myth and History.* In Matthews, M. R., (ed.), *History, Philosophy & New South Wales Science Teaching,Third Annual Conference* (Sydney, 2000), pp. 193-98

Jennings, J. F. (2000). Title I: Its legislative history and its promise. *Phi Delta Kappan*, February, 2000.

Kaestle, C. F. (2001). Federal aid to education since World War II: Purposes and politics. In: *The future of the federal role in elementary and secondary education*. Washington, D.C.: Center for Educational Policy.

Kernan-Schloss, A. (2004). Fighting NCLB's failure label: How to take charge of communicating before the media define your schools as failing. The School Administrator, March, 2004. Retrieved from http://www.aasa.org/publications/sa/2004_03/kernan-schloss.htm

Kim, J. & Sunderman, G. L. (2004). *Does NCLB provide good choices for students in low-performing schools?* Cambridge, MA: The Civil Rights Project at Harvard University.

Kim, J. & Sunderman, G. L. (2004). *Large mandates and limited resources: State response to the No Child Left Behind Act and implications for accountability.* Cambridge, MA: The Civil Rights Project at Harvard University.

Kingsbury, G. & Houser, R. (1997). Using data from a level testing system to change a school district. In *The Rasch tiger ten years later: Using IRT techniques to measure achievement in schools.* Chicago, IL: NATD, 1997.

Kingsbury, G. G., Olson, A., Cronin, J., Hauser, C. & Houser, R. (2003). The state of state standards: Research investigating proficiency levels in fourteen states. Portland, OR: Northwest Evaluation Association.

Lee, J. (2004, April 7). How feasible is adequate yearly progress (AYP)? Simulations of school AYP "uniform averaging" and "safe harbor" under the No Child Left Behind Act. *Education Policy Analysis Archives,* 12(14). Retrieved, April8, 2004 from http://epaa.asu.edu/epaa/v12n14.

Linden, W. van der, & Hambledon, R. (1997). Handbook of Modern Item Response Theory. New York: Springer-Verlag.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, Vol. 32, No. 7, pp. 3–13

Linn, R. L. (2001). *The Design and evaluation of educational assessment and accountability systems.* CSE Technical Report 539, 2001. National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles: UCLA.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). *Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001.* CSE Technical Report 567, 2002. National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles: UCLA.

Marion, S., White, C., Carlson, D., Erpenbach, W.J., Rabinowitz, S., & Sheinker, J. (2002). *Making valid and reliable decisions in determining adequate yearly progress* ASR-CAS Joint Study Group on Adequate Yearly Progress, Council of Chief State School Officers: Washington, D.C.

Natriello, G. & McDill, E. L. (1999). Title I: From funding mechanism to educational program. In The Harvard Civil Rights Project, *Hard Work for Good Schools: Facts Not Fads in Title I Reform.* Cambridge, MA: Harvard University.

Northwest Evaluation Association. (2002). *RIT Scale Norms.* Portland, Ore: Author.

Sanders, W., & Horn, S. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation*, 9, 299-311.

Schugurensky, D. (2002). Elementary and secondary school act, the "War onPoverty" and Title I. In *History of Education: Selected Moments of the 20th Century,* Schugurensky, D., Ed. Retrieved from http://fcis.oise.utotonto.ca/~daniel_schugurensky/assignment1/1965elemsec.html.

Secretary's Commission on Achieving Necessary Skills (SCANS). (1991). *What Work Requires of Schools.* Washington, DC: U.S. Department of Labor.

Seltzer, M., Choi, K., & Thum, Y. M. (2002). *Examining relationships between where students start and how rapidly they progress: Implications for constructing indicators that help illuminate the distribution of achievement within schools.* CSE Technical Report 560, 2002. National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles: UCLA.

Stecher, B. & Arkes, J. (2001). Rewarding schools based on gains: It's all in how you calculate the index and set the target. Washington, D. C.: RAND Corporation.

Thum, Y. M. (2002). *Measuring pregress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis.* A Milken Family Foundation Report.

Thum, Y.M. (2002). *Measuring student and school progress with the California API***,** CSE Technical Report 578, 2002, National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles: UCLA.

Thum, Y. M. (2003). *No Child Left Behind: Methodological challenges & recommendations for measuring adequate yearly progress.* CSE Technical Report 590, 2003. National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles: UCLA.

U. S. Department of Education. (2002). *Title I--Improving the Academic Achievement of the Disadvantaged; Final Rule, December 2, 2002.* Federal Register: December 2, 2002, (Volume 67, Number 231), Rules and Regulations, Page 71709-71771. Washington, D. C.: Author. Retrieved from http://www.ed.gov/legislation/FedRegister/finrule/2002-4/120202a.html.

Young, S. (2004). *History of the federal role in education*. Washington Conference of State Legislatures. Retrieved from: http://www.ncsl.org/proegrams/educ/NCLBHistory.htm.

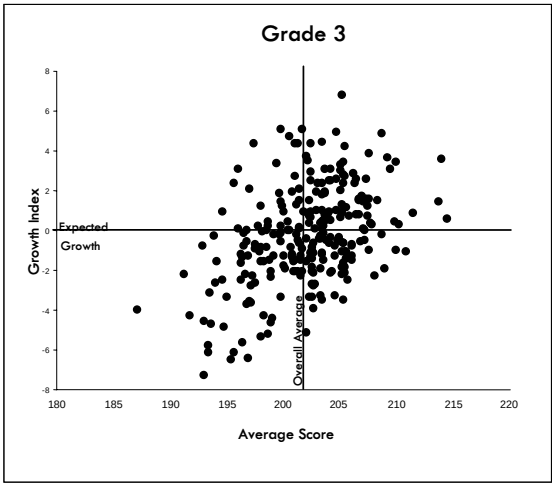
Appendix A

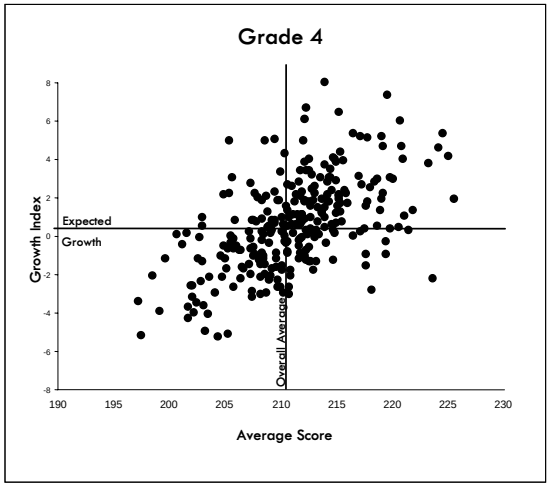
Growth Index and Average Score Quadrants by Grade

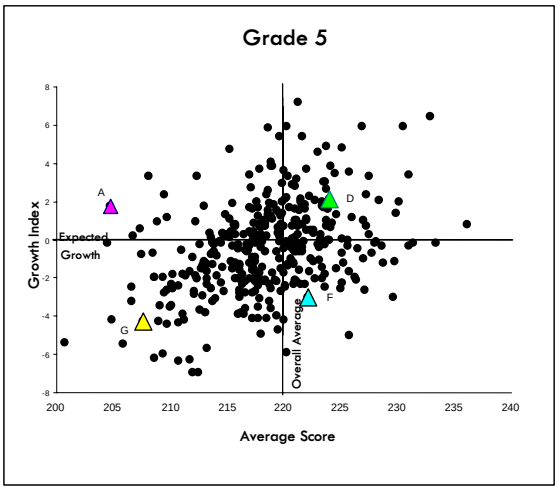
MATHEMATICS

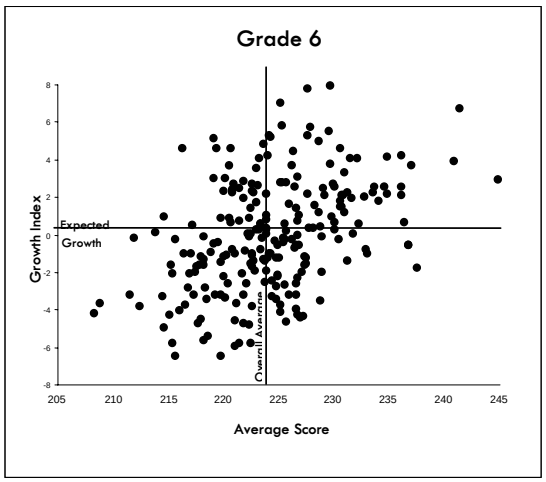
MATHEMATICS							
Grade in 2003	Quadrant	Number of Schools	Students with Sp '03 Scores	Mean Spring 2003 Score	Students with '02 and '03 Scores	Mean Raw Growth	Mean Growth Index
3	Q1	34	2,556	200.0	2,017	14.9	1.8
3	Q2	81	5,998	205.7	5,033	14.3	1.8
3	Q3	64	4,750	204.7	4,007	10.8	-1.5
3	Q4	82	5,997	197.9	4,793	10.7	-2.2
3	Total	261	19,301	202.3	15,850	12.4	-0.2
4	Q1	53	4,056	208.8	3,254	10.9	1.4
4	Q2	111	8,597	215.7	7,086	11.4	2.3
4	Q3	24	1,906	214.5	1,432	8.0	-1.0
4	Q4	87	6,255	206.5	4,914	7.8	-1.7
4	Total	275	20,814	211.5	16,686	9.9	0.7
5	Q1	57	4,277	215.9	3,516	10.6	1.3
5	Q2	94	7,352	223.0	6,077	11.3	2.0
5	Q3	84	7,914	223.2	6,386	7.8	-1.5
5	Q4	139	11,150	214.1	8,798	7.1	-2.2
5	Total	374	30,693	219.8	24,777	8.8	-0.5
6	Q1	46	4,397	221.7	3,512	8.5	1.9
6	Q2	68	7,735	230.2	6,130	9.2	2.4
6	Q3	46	7,426	227.6	5,392	5.0	-1.8
6	Q4	74	12,703	219.4	8,667	3.9	-2.8
6	Total	234	32,261	224.2	23,701	6.2	-0.5
7	Q1	18	2,750	225.7	2,187	8.5	1.4
7	Q2	37	5,800	233.1	4,701	8.8	1.5
7	Q3	52	10,615	233.7	8,074	5.1	-2.2
7	Q4	65	10,959	224.3	8,184	4.4	-2.7
7	Total	172	30,124	229.4	23,146	5.9	-1.3
8	Q1	1	57	226.2	42	8.5	0.1
8	Q2	8	1,021	237.9	894	9.0	0.5
8	Q3	56	8,876	239.7	7,451	5.2	-3.2
8	Q4	53	8,305	230.4	6,456	4.9	-3.5
8	Total	118	18,259	235.3	14,843	5.3	-3.1

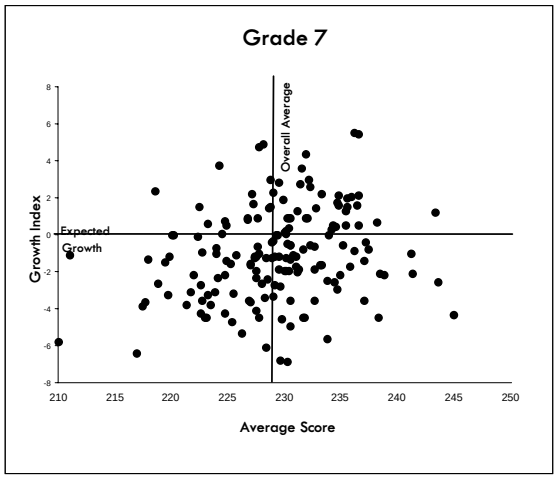
Mathematics – Growth Index by Average Score

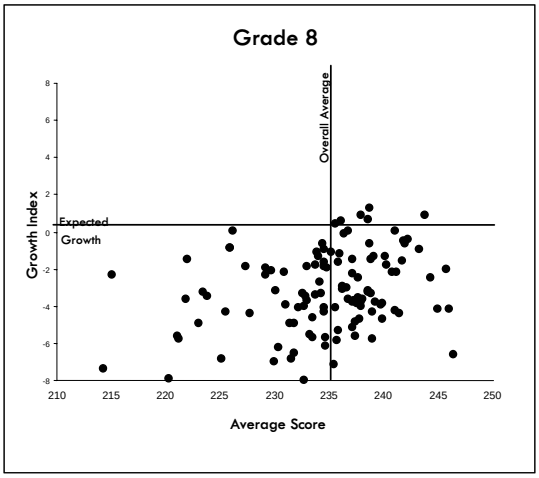












Growth Index and Average Score Quadrants by Grade

READING

READING							
Grade in 2003	Quadrant	Number of Schools	Students with Sp '03 Scores	Mean Spring 2003 Score	Students with '02 and '03 Scores	Mean Raw Growth	Mean Growth Index
3	Q1	21	1,358	193.5	1,060	14.6	1.5
3	Q2	51	4,031	202.1	3,408	12.6	1.1
3	Q3	95	7,155	200.5	5,879	9.4	-1.9
3	Q4	96	7,253	192.4	5,756	9.4	-3.1
3	Total	263	19,797	197.4	16,103	10.4	-1.5
4	Q1	36	2,790	200.4	2,084	10.2	1.3
4	Q2	75	5,986	208.9	4,976	9.0	1.5
4	Q3	70	5,442	207.7	4,466	5.9	-1.3
4	Q4	79	5,925	200.1	4,660	6.5	-2.0
4	Total	260	20,143	204.8	16,186	7.6	-0.3
5	Q1	37	3,195	207.1	2,486	8.7	0.9
5	Q2	69	6,018	214.7	4,800	8.1	1.3
5	Q3	129	11,837	213.6	9,168	5.5	-1.2
5	Q4	131	9,680	206.1	7,587	5.2	-2.3
5	Total	366	30,730	210.8	24,041	6.2	-0.8
6	Q1	26	2,237	211.9	1,669	7.9	1.6
6	Q2	58	6,630	220.5	4,860	6.4	1.1
6	Q3	64	10,647	217.9	7,623	4.1	-1.1
6	Q4	78	12,662	210.4	8,588	3.3	-2.7
6	Total	226	32,176	215.0	22,740	4.6	-1.0
7	Q1	26	4,224	215.6	3,279	6.5	1.3
7	Q2	31	6,093	222.4	4,219	5.3	0.9
7	Q3	49	10,693	221.8	8,164	3.2	-1.1
7	Q4	63	10,282	214.4	7,464	2.9	-2.0
7	Total	169	31,292	218.6	23,126	3.9	-0.7
8	Q1	8	1,160	216.3	840	7.8	2.2
8	Q2	31	5,502	225.3	4,421	5.4	1.0
8	Q3	37	5,798	224.5	4,703	2.9	-1.4
8	Q4	44	7,122	218.7	5,609	2.4	-2.4
8	Total	120	19,582	222.1	15,573	3.7	-0.9

Reading – Growth Index by Average Score