> The primary object of teaching is to produce learning (that is, change), and the amount and kind of learning that occur can be ascertained only by comparing an individual's or a group's status before the learning period with what it is after the learning period.
>
> Frank B. Davis
>
> *Educational Measurements and Their Interpretation*

Focusing State Educational Accountability Systems: Four Methods of Judging School Quality and Progress

Dale Carlson[1]

State educational accountability systems have moved to center stage. Although the stated goal of all these programs is the improvement of learning and instruction, they differ in many significant ways, including the most basic thing of all--their very definitions of school quality and progress. These differences lead to differences in the types of change that schools are rewarded for making, and the types of schools that receive the attention that comes with accountability. Do these different systems flow from fundamentally different beliefs about schools and how they ought to improve? Or do they merely reflect the immature state of a field that, in its youthful vitality, is characterized more by a focus on action and results than on careful attention to arcane definitions and their implications? The purpose of this paper is to help the designers of such systems think more critically and productively about the questions that they want to ask about schools and the kinds of analyses that best answers those questions.

The process begins with the need to answer a question[2]. The question appears in many forms, including the following: How good is this school? Is it getting better? Should this school be identified as in special need of improvement? Should it be given an award for high achievement, or for making outstanding progress?

For this discussion, the first two questions are the most fundamental; they lie at the base of all others:

1. How good is this school?
2. Is it getting better?

These questions probe two very different aspects of a school: (1) its achievement level, and (2) whether that achievement level has changed, that is, its status and its change.

[1] Please send comments and suggestions on this paper to Dalexyz@aol.com.

[2] The comments, suggestions and probing questions of Richard Hill of the National Center for the Improvement of Educational Assessment were immensely helpful in shaping and pruning the ideas presented in this paper.

The two questions can each be subdivided into two additional questions:

1. **How good is this school?**
    a. What is the achievement level of the students?
    b. Is this an effective school? Given the achievement level of students when they enter, how much do they develop or grow while they are in the school?

2. **Is it getting better?**
    a. Is the achievement level of this school improving or declining?
    b. Is the quality or effectiveness of this school improving? How much more or less, are the students learning than they did the year before?

It can be seen that the two main questions are qualitatively different from the sub-questions. The two main questions call for a judgment, one that could easily lead to a decision, possibly followed by some action toward the school. The sub-questions ask factual questions that--at least theoretically--could be categorically answered.

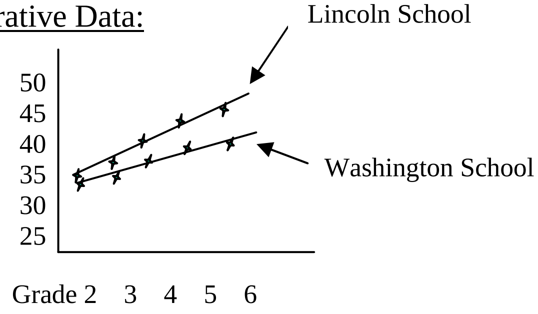
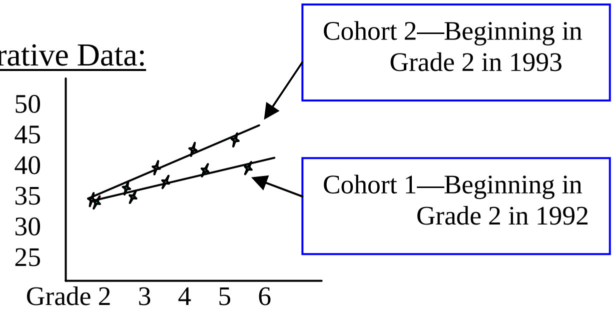
Secondly, it can be seen that the sub-questions are very different from each other. They could easily provide quite different answers to the main questions. The purpose of this paper is to help the reader think about the pros and cons of focusing on each of the main questions, and to think about which of the sub-questions leads to the best answer to the main question.

It may be useful to look at the four questions in a two-by-two matrix, as follows:

	How good is this school? (Status)	**Is it getting better? (Change)**
Achievement	(1a) What is the achievement level of the students in this school?	(2a) Is the achievement level of this school improving?
Effectiveness	(1b) Is this an effective school? Given the achievement level of students when they enter, how much do they learn or develop while they are in the school?	(2b) Is this school becoming *more* effective? How much more, or less, are the students learning this year than they did the year before?

Putting the questions in a matrix highlights the systematic differences between the sub-questions. The first row of the matrix is called “achievement” and the second row is called “effectiveness”. The meaning and importance of this distinction will become more obvious as the arguments unfold (although the reader is encouraged to muster up “a willing suspension of disbelief” at this point!). Exhibit 1 illustrates the meaning of this distinction with some hypothetical data and graphs. This is followed by a more complete discussion of the strengths and weaknesses of each sub-question, and the analytic procedure inherent to that quadrant, in allowing someone to draw a valid conclusion about a school’s quality or progress.

Exhibit 1. Four questions and four types of analyses of school performance data

<table>
<tr><th></th><th>Level of achievement or achievement slope (Status)</th><th>Change in Status or Gradient of Slope (Change)</th></tr>
<tr>
<td><b>Achievement—</b> Average test scores for a given grade and year, and changes from one year to the next</td>
<td><b><i><u>Quadrant A—Question 1a:</u></i></b>
<u>Focus is on status</u><b>: What is the achievement level of students in this school?</b> How well do the students score? How well do they read, write, compute, etc?
<u>Illustrative data:</u>
• Twenty four percent of the third graders in this school can read at the “proficient” level</td>
<td><b><i><u>Quadrant B—Question 2a:</u></i></b>
<u>Focus is on change</u>: <b>Is the achievement level of this school improving?</b> How do this year’s scores compare to last year’s scores—for a given grade level
<u>Illustrative data:</u>
<table>
<tr><th></th><th>1999-00</th><th>2000-01</th><th>Difference</th></tr>
<tr><td>Grade 3</td><td><i>224</i></td><td><i>231</i></td><td><i>+7</i></td></tr>
<tr><td>Grade 4</td><td><i>232</i></td><td><i>225</i></td><td><i>-7</i></td></tr>
<tr><td><b>Average</b></td><td><b><i>228</i></b></td><td><b><i>228</i></b></td><td><b><i>0</i></b></td></tr>
</table>
</td>
</tr>
<tr>
<td><b>Slopes—</b> Progress from one grade to the next, and change in that level of progress</td>
<td><b><i><u>Quadrant C—Question 1b:</u></i></b>
<u>Focus is on effectiveness of the school</u><b>. Is this an effective school?</b> This is indicated by a given cohort of students scoring higher as they move up the grades, reflecting the quality of that school’s program. How well are the fourth graders doing, relative to their performance as third graders? How steep is the slope?
<u>Illustrative Data:</u>
 </td>
<td><b><i><u>Quadrant D—Question 2b:</u></i></b>
<u>Focus in on <i>change</i> in school effectiveness</u><b>. Is this school becoming <i>more</i> effective?</b> This is shown by succeeding cohorts of students scoring <i>increasingly</i> higher, as it reflects increased quality of that school’s program. Is the slope steeper for this year’s cohort? For example, is the gain from second to sixth grade larger for this year’s sixth graders than it was for last year’s sixth graders?
<u>Illustrative Data:</u>
 </td>
</tr>
</table>

<u>Quadrant A--Question 1a</u>: What is the achievement level of the students in this school?

The first cell or quadrant in the matrix focuses on the actual level of achievement of the students. It may be based on a norm-referenced test where the results are reported in terms of national norms, or it may be based on a standards-based assessment where the results are expressed in terms of the percent of students that meet or exceed a given performance standard or performance level.

Exhibit 2 presents some illustrative figures for all the four questions/quadrants. For question #1, the focus would be on the fact that 24% of the students meet the Proficiency level [3]. For the school as a whole, one may want to take the average of the 24% for grade 3 and the 28% for grade 4. Either way, it is a simple performance number, indicative of performance at one point in time.

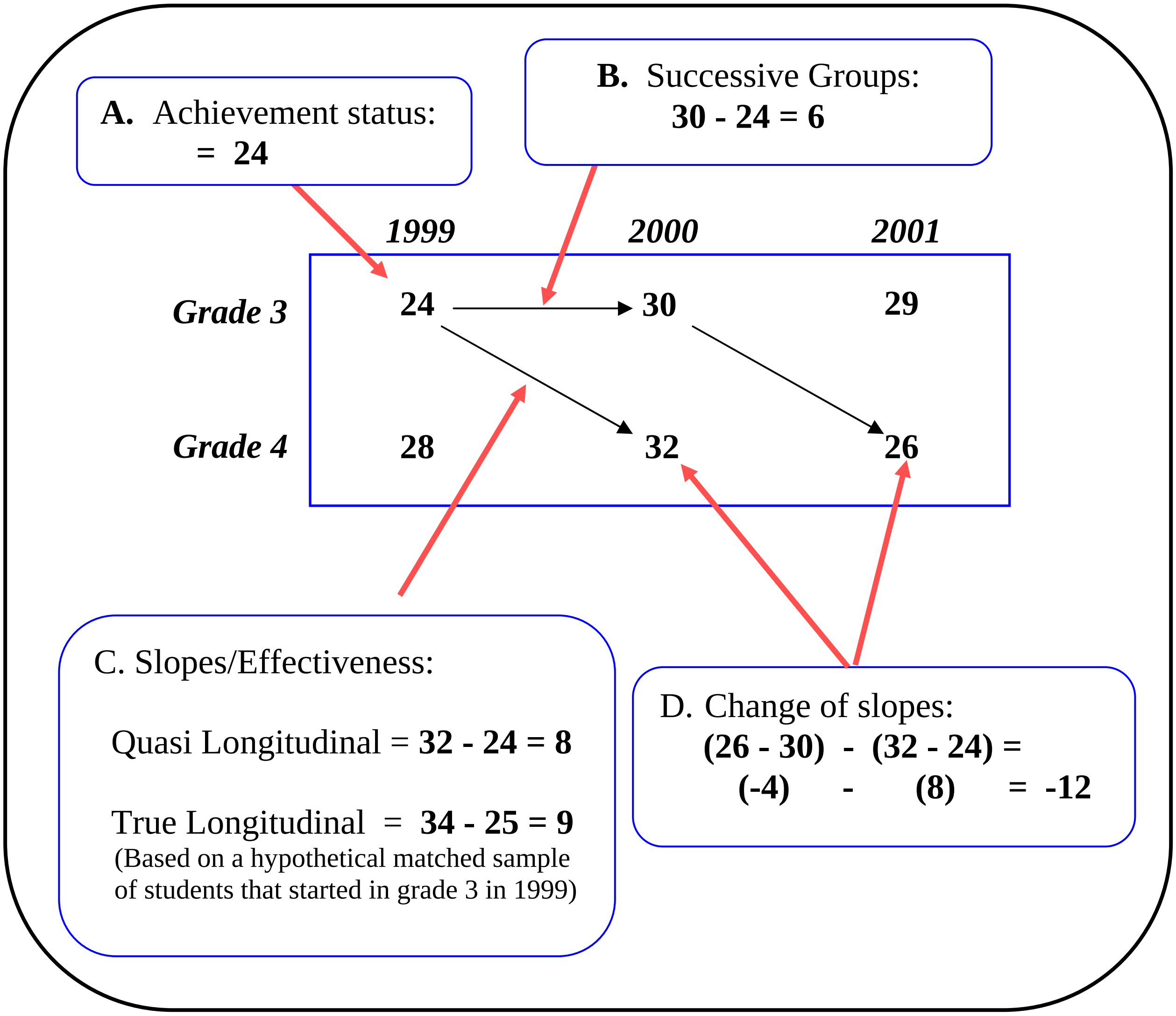
What can be inferred from these kind of results? On one hand, the answer is “much”; on the other hand, the answer is “nothing.” If one wants to make a statement about the actual level of achievement—what the students can do, or how they compare with other students, the answer is clear. If, on the other hand, one wants to judge the quality or effectiveness of the instructional program from the results, the obstacles are severe. The results for a given school with high scores, for example, will almost certainly reflect the quality of instruction, but to an unknown degree the scores will also reflect the nature of the student population that lives in that school’s attendance area. The results are what they are; and they can be useful as long as the inferences are kept in check, perhaps comparable to the level of literal comprehension in reading assessment.

Many statistical strategies have been employed to disentangle the instructional effects from the socioeconomic effects for a given school, with a moderate level of success. This approach is the foundation of one whole school of research on school effectiveness (Sammons, 1999). The problem is that the practice of comparing a school to a set of real or hypothetical similar schools can lead to lowered expectations for schools at the lower end of the scale. A given school with low scores may be doing relatively well when compared to schools with a similarly high level of student poverty. It is virtually impossible, however, to be certain that the recipients of such information won’t—usually inadvertently—draw the conclusion that the school is, therefore, doing about as well as expected. The power of expectations to improve the level of learning in a school or to retard that level of learning is no longer open to debate. Proponents of standards-based reform are rightly skeptical about the use of such devices.

Can the type of information used to answer question 1a be the basis for setting goals or defining adequate yearly progress? One large State (Texas) does exactly that. Schools are classified on a four-level rating system according to the percent of students that pass the reading and math tests at each grade level (as well as other indicators of school quality at some grade levels). It is a simple approach, but that simplicity pays off in ease of use and ease in communicating to the public.

[3] This assumes that the performance levels have a common meaning and symmetry across the grade levels. For comparing the relative status of different schools, however, a reasonable amount of asymmetry may be tolerable as long as it applies equally to all schools.

Exhibit 2. A Numerical Illustration of the Four Quadrants



A. Achievement status:
= **24**

B. Successive Groups:
30 - 24 = 6

	1999	*2000*	*2001*
Grade 3	**24**	**30**	**29**
Grade 4	**28**	**32**	**26**

C. Slopes/Effectiveness:

Quasi Longitudinal = **32 - 24 = 8**

True Longitudinal = **34 - 25 = 9**
(Based on a hypothetical matched sample of students that started in grade 3 in 1999)

D. Change of slopes:
(26 - 30) - (32 - 24) =
(-4) - (8) = -12

Quadrant B—Question 2a: Is the achievement level of this school improving?

The second question focuses on change. It looks at the difference in achievement from one year to the next, and does so for successive groups at a grade level. For example, it compares the results for grade four for 1999-2000 with the scores for a new group of fourth graders in 2000-2001. Looking at Exhibit 1, that would represent the difference between the scores of 231 and 224, leading to a gain of 7 for grade 3.

This is the dominant design for studying progress of schools across the nation. The problem is an obvious one: it assumes that one can infer changes in the quality of a school or its programs by looking at the difference between two different groups of students at different points in time. The example on Exhibit 1 illustrates a typical finding. If one looked only at the results for grade three, the conclusion would be that the school was improving. If one looked only at the results for grade four, the conclusion would be that the school was declining; and if one looked at the average of both, the conclusion would be that the school was constant. What is the real level of improvement?

Two factors work together to mitigate the meaningfulness of successive group differences: initial group differences and on-going, year-to-year mobility. All teachers know about the large and random differences between successive classes. Their hunches about the "good class-bad class" phenomena are verifiable, and in fact, are larger than most people assume. Then there is the on-going mobility problem. The key issue is the relative difference in achievement between the students moving out and those moving in—obviously! The surprise is how unpredictable these differences are, and how strongly they can effect the validity of inferences for large schools or schools with relatively stable populations. So considering both obstacles, the difficulty of linking test score changes to changes in the quality of instruction is nearly impossible.

It is nearly unbelievable how little attention has been given to the validity of this approach, especially considering its widespread use. Many people assume that the mobility problem exists but that it isn't much of a problem for large schools or for schools with low mobility rates--or that it washes out with several years of data. There is some truth in all of these assumptions. What research is showing, however, is that it takes at least three or four years of data to draw a valid conclusion; secondly, that large schools are not immune to the effects of mobility or initial differences; and, thirdly, that surprisingly low levels of mobility can render year-to-year, successive-group differences virtually uninterpretable.

This approach makes sense if the evaluation process has a long timeframe. The success of schools that are truly improving will be reflected in the continuously improving scores of different cohorts of students. As each new cohort is the beneficiary of a continuously improving program, its scores will be higher than the preceding cohort (measures at some mid-point or end-point, obviously not at the beginning grades in a school). In the short term—the timeframe for most accountability systems, however, the results are not encouraging. The findings from this approach often lead to very different judgments of schools than if one looked just at the progress of the students who had the benefit of a school's program—the focus of quadrant C (Carlson, 2001; Hill, 2001). Linn and his colleagues found the same trends long ago (Dyer et.al., 1969), and warned researchers of the need to take precautions to avoid drawing false conclusions on the basis of successive-group differences.

Quadrant C—Question 1b: Is this an effective school?

If student learning is the chief business of schools, then the amount that students learn—as reflected in the progress they make from one grade to the next—should be the best measure of a school's effectiveness or quality[4]. Obviously, students will start at different points or levels, but what counts is the growth they make. Rogosa (1982) set forth the basic argument that individual growth curves are the most logical units for studying a student's progress. By extension, the aggregation of individual growth curves then is the most meaningful reflection of a school's impact. New methodologies are beginning to be applied to the study of these aggregate curves; methods which yield productivity indices for schools (Bryk, 1998) or indicate whether interventions have been effective (Bloom, 2001). It is hoped that this paper will foster discussion and research on the application of this slopes-as-outcomes approach to school accountability.

Schools with steep growth curves are considered more effective and schools with flatter growth curves less effective. As shown in the illustration on Exhibit 1, if Lincoln School moved its students from the 35th percentile at the end of grade three to the 40th percentile rank at the end of grade four, that school is considered to be less effective than Washington School which moved its students from the 34th to the 49th percentiles[5]. The example in Exhibit 2 shows a schoolwide growth rate of 8%, a change from 24% of the students meeting the standard at the end of grade three to 32% of the students meeting the standard when they reached the end of fourth grade.

Since this is essentially a longitudinal approach, the user has a choice of whether to use a matched or a non-matched longitudinal sample. A matched longitudinal approach would use only the scores for the students who were present at both assessments. The comparison could also be based on all the students who happened to be in grade three in 1999, however, and all the students who happened to be in the school in grade four the next year, an unmatched longitudinal sample. The former has been labeled a true-longitudinal comparison and the latter a quasi-longitudinal comparison (Linn, 2001). The figures in Exhibit 2 illustrate both approaches. The quasi-longitudinal numbers for the school are the same numbers (percent of students reaching the proficient level) used to answer questions 1a and 2a. The hypothetical figures for the true-longitudinal approach assume that the stable students were a little higher at the end of grade three (25% versus 24% for the quasi-

[4] Quality is used synonymously with effectiveness in this paper, knowing that the judgment of a school's true quality usually includes many factors that are not tapped by most assessment systems.

[5] This is not to say that the choice of metric is irrelevant, or that the assessments used at different grade levels wouldn't be required to demonstrate certain content and measurement properties. It is suggested here, however, that for certain limited purposes related to the identification of effective or ineffective schools, these requirements may be quite minimal. Furthermore, the more widely used a common set of assessment instruments, the broader would be the realm of inference. That is, if different school districts used very different assessments, one could not be sure that a school that is judged effective in one district would be judged effective in others. Since all states must have common assessment systems for Title I schools, or for all schools, this is not viewed as an insuperable problem. The interpretation of the growth or difference scores, however, is subject to all the statistical complications of change or difference scores, including the intricacies of the relationship between initial score levels and growth scores. Although they are not discussed here, this paper does not dismiss these problems; the emphasis here is on the conceptual advantages and disadvantages of each of the approaches, that is, the proper match between the question that is or should be asked and the analytic approach employed. Furthermore, the actual computational procedures required to implement these approaches may be considerably, if not profoundly, more complex than the simple calculation of differences used to illustrate the four main questions discussed here.

longitudinal approach) and made very slightly more progress to reach a higher percent (34% versus 32%) by the end of grade four.

What are the advantages and disadvantages of the two different approaches? Obviously, the true longitudinal approach allows for a less clouded interpretation, since the comparison is not obstructed by the scores of students who left during the year, or entered during the year. On the other hand, the difficulties in obtaining matched samples is profoundly more difficult—virtually impossible if the State does not have a computerized student data system. Fortunately, current investigations are hinting at a most felicitous finding: the results (year-to-year slopes or differences) for the two longitudinal methods are very similar for many schools. A high correlation between the slopes for a set of schools does not necessarily mean that the quasi-longitudinal results accurately reflect that school's true effectiveness—for a given school. Obviously, the comparability of the results for the two approaches is related to the mobility level of the schools. It may be possible to develop procedures that a school can use to check on the validity of the quasi-longitudinal findings, and, therefore, the need to use matched longitudinal samples. With the judicious use of this approach, it may be possible to estimate the true effectiveness of most schools, at least to do so in a manner that is sufficient until student data systems are implemented.

The question and approach described in Quadrant B, "Are the achievement scores changing for this school?", is the approach selected by virtually all state accountability systems, yet the approach described in Quadrant C actually focuses more directly on school effectiveness. This begs the question, "What is the relationship between the results from the two approaches?" Do schools with strong positive changes for successive groups also have positive growth--steep slopes? Given the level of student mobility in American schools, the answer can be anticipated: initial studies show a very weak relationship among the methods (Dyer, et.al., 1969; Carlson, 2001; Hill, 2001). The correlation between the two approaches for several large samples of schools range from 0 to .50, depending on the number of content areas, grades and years involved. Results also varied according to level of student mobility, as expected; however, the extreme sensitivity of this relationship to mobility was surprising. Even for schools with relatively stable populations, the successive group differences did not match the longitudinal findings. Cross-tabulations showed that for a group of schools that might be selected for school improvement based on decreasing successive-group scores, for example, as many as one-fourth of them would be considered extremely effective schools looking at their grade-to-grade slopes!

Is it logical to look at longitudinal differences? It may seem unrealistic to judge a school on the basis of just the students who stay for at least a year in the school; they might be a very selective group. A school might be very effective with the stable students, at the expense of the new students—or vice versa. The problem is that unless students have been in a school, at least for a few months, there is no other basis for judging its effectiveness. The reader needs to understand that the difference between two scores for a school consists of two components (to oversimplify just a bit): (a) the effectiveness of the school in helping students learn, and (b) the combined impact of initial differences between classes or cohorts, and on-going changes to the cohorts due to mobility. That's all there is! So if the impact of "b" is significantly larger than "a", it is easy to see why question 2a (quadrant B) is a weak design.

This leads to the issue of the purpose of the schools and the focus of an accountability system. It is clear that the schools are responsible for educating all the children they enroll. The question is, are there accountability methods that allow their effectiveness with all children to be assessed? It is hard to be optimistic about the answer. Both questions 1a and 2a focus on all children; unfortunately, they do not offer much help in judging the effectiveness of the schools' programs. It might be best to think of the answers to these two questions as indicative of the level of "need" in a school (1a), and the changes in that level of need (2a). As indicators of need, the answers could still play a strong role in an accountability system, both in focusing efforts to improve the programs, and moderating the decisions about improvement and rewards. For example, a State may decide to give a larger portion of reward funds to schools that started at a lower baseline (1a), or to schools where the level of need is increasing (2a).

Would a quadrant C, question 1b focus act as a perverse incentive? Would schools attend equally to the needs of all children if they knew that the payoff was based on the progress of more stable students? It is hard to see how this could happen since they don't know which students will move out, or how they will compare with incoming students. Others worry about the incentive for schools to generate false progress by doing whatever they can to lower the baseline. If, for example, the focus is on growth from the end of grade three to the end of grade four, what incentive is there for the schools to be sure that students do well on the third grade assessment. This is surely an issue worthy of consideration. It reflects the need to think hard about the design of the overall accountability system, including the interaction among the parts.

All of this serves to underscore the premise of this paper: it's important to ask the right question, and it's equally important to select the analytic method that actually answers that question.

Quadrant D--Question 2b: Is this school becoming more effective?

The matrix in Exhibit 1 deals with both status and change. The top two quadrants, A and B, illustrate status and change as they apply to a single point of achievement (e.g. grade 4 reading scores for 2000) and changes in that single point (the difference between grade 4 reading scores in 2000 and in 2001). The two bottom quadrants, C and D, deal with status and change as they apply to a trend line, or a growth curve across grade levels (akin to the average of the student-level growth curves). For quadrant C, the line is still a static measurement; since it is a function it just takes more than one year's data to construct it. In quadrant D, the focus is on *changes* in that line. Is it flatter or steeper than it was at some previous point for a school? This puts the spotlight on change in a school's effectiveness; if the line is getting steeper, the inference is that the school is becoming more productive or effective.

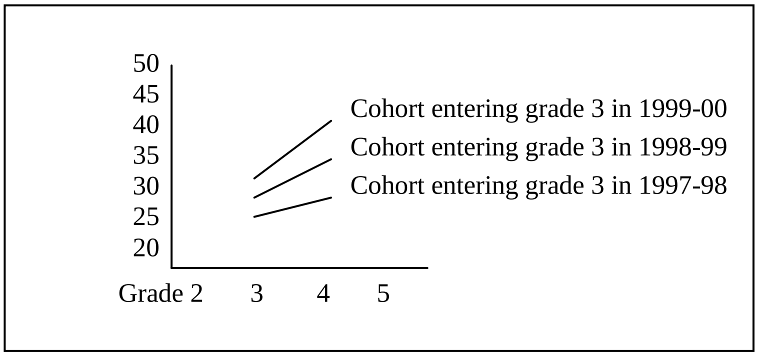
The example on Exhibit 1 shows the growth curves for two cohorts entering a school in two consecutive years. It can be seen that they both start at about the same level, but one rises faster than the other. The focus here is on the difference between the growth curves. Looking at the illustration in Exhibit 2, it can be seen that this school had a strong growth of 8% from third to fourth grade for the first slope, but then the growth for the second line actually shows a decline of 4% from grade three to grade four.

One of the truisms about any kind of achievement score data (like all measurements) is that status scores are always more reliable than difference scores. It could be shown that the most problematic of all would be the differences computed to answer question 2b. This is an unfortunate fact (with parallels in other aspects of measurement--and life): the things we are most interested in are the most difficult to measure. The concept of AYP is a question 2b endeavor. If a school uses a typical gap-reduction approach as its AYP definition, saying, for example, that a school must reduce its percent of students scoring below Proficient by 10% a year, it can be shown that the school must increase its effectiveness or productivity as defined in question 2b. This means, ironically, that the most interesting inference to make about a school: “Is it making AYP?” is the most difficult to make with the level of certainty that should accompany rewards or sanctions.

Short-term longitudinal approaches

The visual display in Exhibit 1 shows growth over several grades, but an assessment system that supplies contiguous, multi-grade data is not essential to judge effectiveness. Exhibit 3 illustrates the use of data for only two contiguous grades to draw inferences about a school’s effectiveness. More grades over more years are obviously preferable; the broader the foundation, the more generalizable the inferences. To the degree that the process of judging school progress is based on aggregated student-level growth curves, it is clear that a multi-year, multi-wave approach is clearly superior, allowing much stronger inferences about a student’s or a school’s progress (Willetts, 1989).

Exhibit 3. An illustration of a short-term longitudinal model



Implications of the models for accountability systems

The four approaches are ways of looking at student achievement and drawing inferences about a school. In some ways, all four approaches are asking about the quality of learning in a school, and the impact or contribution of the school in fostering that learning. Only the two approaches that focus on change (quadrants B and D) could be said to focus on improvement in that level of learning.

The four approaches are not accountability models, but could be used, perhaps in some combined form, to construct a definition of adequate yearly progress that would classify schools for rewards or special assistance. Exhibit 4 sets forth some of the assumptions of each of the approaches; it then proposes a typical AYP definition for each approach; then lists some strengths and weaknesses of each.

Exhibit 4. Some implications of the four questions/approaches for use in accountability systems

	Assumptions and implications for an accountability system	**Illustrative AYP definitions to match the approaches**	**Issues and challenges**
A. Achievement: Status	- Low-performing schools are the main problem in American education; these schools need to reach a minimal level of performance	- All schools with more than 50% of their students scoring below Basic are identified for School Improvement	- This approach identifies, without differentiation, two types of schools: those with a large proportion of economically disadvantaged students, and those with less effective programs.
B. Change in achievement: Successive groups	- Virtually all schools need to improve (regardless of how effective they may be with students who don't move); lower performing schools may need to improve more rapidly	- All schools need to make 5% gain each year, or - All schools need to close the gap between current performance and the goal, e.g., 80% of students reaching the proficient level in 12 years	- Random differences between incoming classes and mobility effects seriously distort the selection of schools that are improving or not improving. - Change scores are considerably less reliable than status measurements
C. Effectiveness: Grade-to-grade slopes	- Virtually all schools need to be more effective; those which are especially ineffective (or are both low in initial achievement and are relatively ineffective) may need to reach a minimal level of effectiveness	- All schools with a lower percentage of students Proficient at grade 4 than the previous year at grade 3 must make those percentages equal within two years, or - All schools with less than 20 % of their students Proficient and a slope less than 8 points must achieve a slope of at least 8 points within two years.	- Computing results for matched-longitudinal samples is difficult logistically - Quasi-longitudinal results may not be accurate for a given school - Change scores are less reliable than status scores, and slopes are change scores
D. Change in effectiveness: Changes in gradient of slopes	- Virtually all schools need to be increasing their effectiveness; those which are making less progress in raising their effectiveness rate (and are low in initial achievement) may need to improve at a higher rate.	- All schools with more than 50% of their students below the basic level in grade 3, and have an effectiveness rate of less than 5% between grades 3 and 4 must increase their rate by 3% both of the next two years.	- Change scores are less reliable than status scores, and these scores are even less reliable since they are differences between change scores.

It can be seen that several of the AYP definitions use a combination of two approaches. Typically they use the general level of achievement (Quadrant A) and one of the others. This is consistent with Title I law and intent, and with many of the approaches taken by states, that is, the intention is to focus more and low-performing schools and to set higher expectations for them to improve.

New directions and next steps

What are the implications of these ideas for States as they define their accountability systems? States are obligated to think hard about the specific goals of their systems and to be sure the approach selected matches their goals. This is not easy. It may even require some analysis of information about their own schools, including the level and nature of student mobility, and the probable impact of using different methods. They need to work with other states to share findings and work on common problems, and they need to work with researchers who are studying these “real world” problems.

What are the implications for researchers? Many questions remain about the different approaches and the ways in which states can draw the wrong conclusions about schools—usually without knowing it. The relationship among the approaches needs to be much more thoroughly studied. Which types of schools are judged differently by the different approaches, and how large are those differences? The stability or reliability of the different approaches is extremely important. Only with careful analysis of large data sets will it be possible to provide guidance to states about the advantages and disadvantages of the different approaches. States will be making decisions about validity-reliability tradeoffs; they need to have some general guidelines for those decisions. New statistical approaches, such as latent growth curve methodology, are becoming more widespread (Bryk, et al. 1998; Thum, in press). This work is currently in the context of individual students; the study of the applicability of these to the study of school effectiveness is urgent.

References

Bloom, H. S. (2001). *Measuring the impacts of whole school reforms: Methodological lessons from an evaluation of accelerated schools.* Manpower Demonstration Research Corporation. Washington, D.C.

Bryk, A. S., Thum, Y. M., Easton, J. Q., and Luppescu, S. (1998). Assessing school academic productivity: The case of Chicago school reform. *Social Psychology of Education*. 2, 103-142.

Carlson, D. C. (2001). *Using cross-sectional designs to evaluate schools for Title I: Roulette anyone?* Paper presented at the American Education Research Association conference, Seattle.

Collins, L. M., (1991). Measurement in longitudinal research. In Collins, L. M. and Horn, J. L. (pp 137-148). *Best methods of the analysis of change: Recent advances, unanswered questions, future directions*. American Psychological Association. Washington, D.C.

Collins L. M. and Sayer, A. G. (editors), (2001). *New methods for the analysis of change.* American Psychological Association. Washington, D.C.

Dyer, H., Linn, R. L., and Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. *American Educational Research Journal*. 6 (4). 591 – 605.

Fitz-Gibbon, C.T., and Morris, L. L. (1987). *How to design a program evaluation*. Newbury Park: Sage Publications.

Hilton, T. L. and Patrick, C. (1970). Cross-sectional versus longitudinal data: an empirical comparison of mean differences in academic growth, *Journal of Educational Measurement*. 7. 15-24.

Kane, T.J., and Staiger, D.O. (August, 2001). *Volatility in school test scores: Implications for test-based accountability systems.* www.brook.edu/gs/brown/brownhp.tm

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems* (CSE Tech. Rep. No. 539). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Mirikitani, R**.**T. (1991). Explaining change: The difference between longitudinal and cross-sectional models of school effects. *New Zealand Journal of Educational Studies.* 26 (2). 125-142.

Muthen, B. O. and Curran, P.J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*. 2. 371-402.

Rogosa, D., Brandt, D., and Zimowski, M.F. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*. 90.726-748.

Sammons, P. (1999). *School effectiveness: Coming of age in the twenty-first century.* The Netherlands: Swets and Zeitlinger.

Stanley, J. C. (1985). Historical note about cross-sectional versus longitudinal studies. *Journal of Special Education.* 19(3). 359-362.

Thum, Y. M. (2001). Designing school performance and school productivity indicators. (CRESST paper, 2001)

Willetts, John B. (1989). Questions and answers in the measurement of change. *Review of Research in Education*. (15) 45-422

Willms, J. D. (1992). *Monitoring school performance: A guide for educators.* Washington, DC: The Falmer Press