Center for
K–12 Assessment
& Performance Management

An independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Exploratory Seminar:

Measurement Challenges Within the Race to the Top Agenda

December 2009

Student Growth Data for Productivity Indicator Systems

Edward Haertel

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Student Growth Data for Productivity Indicator Systems

Edward Haertel

Stanford University, Palo Alto, California

This paper was presented by Edward Haertel at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download copies of other papers presented at the seminar at http://www.k12center.org/publications.html.

Foremost among its intended functions, formal schooling is to equip students with knowledge, skills, and abilities necessary for informed citizenship, personal fulfillment, and successful work force participation. By and large, children are far more capable at the end of their formal schooling than at the beginning. Most have learned to read and compute and have attained some degree of cultural awareness and scientific literacy. Many leave school with habits and dispositions that will serve them well as lifelong learners. It is glaringly obvious, however, that not all students are equally well served by our educational system: Some leave high school well prepared for success in our most elite and demanding colleges and universities; others fail to secure even a high school diploma, or remain functionally illiterate after 12 or more years of schooling. It is clear that teachers, schools, and school systems are profoundly unequal in the quality of education they provide, but it is also clear that some determinants of student achievement are beyond the control of the educational system. Achievement is affected not only by formal schooling, but also by individual aptitude and effort, by peer culture, and by home and community resources, including the alignment of students' home cultures with the implicit and explicit expectations and demands of classroom learning environments.

These two realities—that teachers, schools, and school systems are unequal in their effectiveness and that student learning outcomes are only partially determined by formal schooling—frame the problem addressed in this paper. Unequal educational effectiveness brings a pressing policy need for metrics to evaluate and compare teachers, schools, and districts. There is an understandable desire to identify low performers for heightened scrutiny or intervention and to identify high performers so as to discover the keys to their success. How, then, are teachers and schools to be compared? An obvious, direct, and face-valid index of effectiveness is student achievement. Unfortunately the second reality cited, that achievement is influenced by multiple factors both within and beyond the classroom, greatly complicates the use of achievement outcomes for evaluation of schools or school personnel. Simple comparisons of average test scores for different teachers or schools are rightly regarded as unfair, because different student populations bring different kinds and degrees of educational challenges. A highly effective school serving low-income English learners whose parents have little formal schooling may have lower test scores than a mediocre school serving a more advantaged student population.

The past decade has brought increasing attention to *student growth measures* as alternative indexes of teacher or school effectiveness. These measurements are designed to quantify the *change* in student

achievement over some defined period of time. The logic of this approach holds that even if students begin at different starting points, teachers or schools may be fairly evaluated on the basis of the *increment* in student achievement realized over the course of an academic year or longer. The use of growth measures to supplement or supplant *status measures*—indicators of achievement at a single point in time—can significantly improve the validity and fairness of educational accountability systems and may also improve their reliability. Viewed in this light, as a substantial if incremental improvement upon the use of status measures, student growth measures hold considerable promise. Unfortunately, as with many innovations in education, growth measures have been oversimplified and oversold.

This paper considers the use of student growth measures in indicator systems for productivity analysis. It will be argued that much thought and care must go into the design of any educational indicator system and that different applications will call for different design choices. Growth measures may be useful building blocks in such systems, but no simple formula exists for their proper use or interpretation. In particular, the use of growth measures cannot finesse the tension between learning expectations framed in terms of improvement versus attainment. We may applaud exceptional improvement, but what matters ultimately is the final achievement level attained. Following this introduction, the first major section takes up the problem of measuring growth for individual students and recasts growth measures in different terms more amenable to analysis; the second section considers the aggregation of individual measures to characterize the status of groups (e.g., the students in a classroom, a school, or some demographic subgroup within a school); and the third considers the use of such group performance measures as part of educational indicator systems for productivity analysis. The paper ends with a brief summary and conclusion.

Defining Growth for an Individual Student

The phrase *student growth* seems clear when considering change in achievement from some starting point in time to some end point. Although the term is used throughout this paper, it may be a bit misleading in the context of models incorporating test scores from multiple prior points in time. In this section, the problem is framed not as one of constructing a growth measure per se, but instead of adjusting end point test scores so as to account for those achievement influences that are irrelevant to the intended use of the end point test scores. Consider the problem of comparing the effectiveness of teachers or of schools using their students' test scores. Simply comparing end-of-year (end point) test score averages is intuitively appealing, but those average scores are unsatisfactory because they will in part reflect variables that are outside the control of the teachers or schools being compared. If it were possible to sort students into groups that were homogeneous with respect to those external factors, then within each such group, variation in end-of-year test score means could be attributed to the relative effectiveness of the teachers or schools responsible for the students' education. Stated differently, rather than comparing final mean scores for all students, a more refined approach would be to compare the relative effectiveness of teachers or schools in educating students who were *similarly situated with respect to external factors*. If external factors could be held constant in this way, then (after accounting for sampling and measurement error) teachers or schools might be held accountable for the remaining score variation.

One could generate an endless list of potential achievement influences beyond the control of schools or teachers, but growth measures hold out the promise of a simpler accounting for all such external factors. It may be argued that whatever those external influences might be, they should be reflected in students' achievement test scores from the prior year (or the beginning of the relevant period of instruction). If so, then an intuitively appealing approach is simply to group students according to their prior year test scores and make comparisons among students who began at the same starting point. Even better statistical control for preexisting student differences could be attained by using test scores from two or more earlier points in time and grouping students according to their score profiles across two or more prior years.[1] This rationale clarifies the construct that student growth measures are intended to capture. From the perspective of educational evaluation or productivity analysis, the purpose of these measures is to disentangle those influences on student achievement for which teachers or schools are rightly held accountable from those that lie beyond their control. If that separation could somehow be accomplished, then comparing the relative effectiveness of teachers or schools would be straightforward.

Even if students could be sorted into similarly situated groups—groups that were homogeneous with respect to external factors—it would be impractical to conduct separate teacher or school comparisons for each such group. Instead, a statistical index of achievement relative to initial status is required. Ideally, that index should be interpretable in the same way for all students, regardless of their prior achievement patterns, so that the average for a group of students is meaningful. If just two scores are available for each student—a pretest and a posttest—and if these share a common (vertical) equal-interval scale, then the most obvious and perhaps best index is the gain score (defined as posttest minus pretest). Growth in reading or mathematics proficiency is thought of as movement along a line, and the gain score, a measure of the distance from the starting point to the end point, or *value added*, is taken to provide a measure of teacher or school effectiveness that is more accurate and more fair than a measure reflecting only the end point reached without regard for the starting point. Alternatively, with one or more prior years of data available, and with or without a vertical scale, an index might be created by replacing observed end-of-year scores with values adjusted for prior achievement patterns using a statistical regression model. Teacher or school comparisons are then based on the means of these adjusted values (residualized gain scores). Either gain scores or residualized gain scores represent end point achievement adjusted for initial achievement level or for achievement levels at two or more earlier points in time. However, it must not be forgotten that comparisons based on these student growth measures are only fair and accurate to the extent that patterns of prior test scores in fact capture all those achievement influences outside of teachers' or schools' control.

Formally, the growth measure is a scalar variable (i.e., a single numerical value for each student) defined as a function of some collection of measurements (a vector). At a minimum, that vector includes one or more achievement scores obtained at the end point, as well as at least one achievement score from some earlier point in time, ideally close to the beginning of the period of instruction over which teachers' or schools' performances are to be compared. In the most general formulation, the score

[1] This is where the language of student growth becomes a little confusing, or even misleading.

vector might include observations from several earlier time points, and each such observation might include multiple test scores as well as nontest information. A mathematical function maps that vector-valued observation into a scalar index intended to represent the student's end point achievement net of external influences. This general definition must be made more precise for any actual measurement application. A precise definition can then guide the optimum use of available data to construct a growth measure and can also highlight any remaining defects of imprecision, construct underrepresentation, or construct-irrelevant variance. The interpretation of these adjusted scores will in general be norm-referenced. There is no need to specify any specific amount of growth expected.

For the simplest case where only a pretest and a posttest are available, the simple gain score was described earlier. Another index that might be used is the *student growth percentile* (SGP). One advantage of the SGP is that it does not require a vertical scale. Each student's SGP is the percentile rank of her or his posttest score within the distribution of posttest scores for all students sharing his or her pretest score. In other words, students are grouped by pretest, and posttest scores are then converted to percentile ranks within each pretest score group. If a student's posttest score is at the median of posttest scores for students sharing the same pretest score, then that student's SGP is 50. This kind of regression model makes minimal assumptions. The SGPs are invariant under any monotone transformations of the pretest and posttest score scales. Betebenner (2009) described and illustrated the advantages of this kind of norm-referenced growth interpretation.[2] If the only prior information available concerning each student is that student's pretest score, then the definition of *similarly situated students* as those with the same pretest may be the best available. However, at least three difficulties arise in connection with the assumption that posttest scores of students sharing the same pretest score should be directly comparable. These same difficulties arise if multiple prior test scores are available.

Problems in Assuming That Students With the Same Pretest Score Are Similarly Situated

The first problem in defining a common posttest expectation for students with the same pretest score is measurement error. The observed score at pretest may be regarded as the sum of a true-score component and an error component (measurement error). Ideally, students would be grouped according to their pretest *true* scores, not their pretest *observed* scores. The second problem is that students with the same pretest observed score but differing with respect to other characteristics may on average have different true scores. This instance belies the familiar statistical phenomenon of *regression toward the mean*. For example, consider two groups of students with the same pretest score: one group from households with low socio-economic status (SES) and the other from high-SES households. Assuming that the low-SES group performs more poorly overall on the pretest, the average pretest true score for the low-SES group will be lower than for the high-SES group *even when comparing subgroups with the same observed pretest score*. The third problem raises larger questions about the adequacy of

[2] The definition of SGPs is readily extended to include data from two or more prior time points. Accuracy improves as additional waves of data are incorporated, but all of the issues raised in this paper still pertain to a greater or lesser degree.

statistical control for prior achievement. Even if pretest true scores could be known without error (i.e., even if pretest achievement could be perfectly measured), it might still be inappropriate to hold a common posttest achievement expectation for all students with the same true pretest achievement level. These three concerns challenge the adequacy of any growth model, not just SGPs.

Outside Influences on Achievement Not Captured by Prior Test Scores

To see why pretest achievement, even perfectly measured, might be inadequate as a basis for sorting out factors within versus beyond educators' control, three scenarios are considered. The first highlights differences in individual aptitude, the second highlights differences in out-of-school supports for learning, and the third highlights differences in the quality of schooling itself. Each scenario is framed in a context where one particular group of factors is especially salient, but additional complexities are also noted.

The First Scenario

Consider the problem of comparing elementary school teachers with self-contained classrooms on the basis of their students' growth in reading since the previous year. Suppose that some teachers are assigned to gifted and talented education (GATE) classes, some to sheltered English classes (for English learners), and some to regular classes. Within the regular classroom group, teachers have anywhere from zero to four mainstreamed special education students in their classrooms. Imagine that a perfectly reliable (i.e., error-free) reading pretest was available, and that these ideal pretest measures were used, together with students' posttest scores, to calculate their SGPs. Each teacher's performance was then quantified by the median SGP for his or her classroom. Direct comparison of teachers on the basis of these growth outcome summaries would be widely regarded as unfair because of differences in student aptitudes. ("Aptitude" is used broadly here to refer to all factors determining a pupil's readiness to profit from instruction. This definition encompasses much more than just native ability.) A younger, precocious GATE student and a somewhat older student in a regular classroom might have the same pretest score, but other things being equal, the GATE student might still be expected to progress more rapidly over the interval from pretesting to posttesting.

The complication of noting that teachers have varying numbers of special needs students highlights the additional problem that students within a classroom are not independent of one another, and the presence of students who are disruptive or who place greater time demands on teachers may on average affect the performance of other students as well.

Still more problems may arise due to limitations of the posttest measure used. As noted, the SGP is invariant under monotone transformations of the score scale. Unlike a simple difference score, it does not require any assumption of an equal-interval scale. Even with this metric-invariant statistic, however, measurement error can introduce bias. Most achievement tests are designed to be somewhat more accurate in the middle part of their score ranges than at the extremes, because by concentrating most of the test's discriminating power in the region of the score scale where most students are located, overall accuracy is improved. But this means that on average, posttest scores and the SGPs derived from

them are likely to be less reliable for the GATE and sheltered English students than for those in regular classrooms. For the pretest, this problem has been assumed away by invoking some ideal, perfectly reliable test, but the problem remains for the posttest. Lower reliability at the extremes will work to the advantage of below average GATE and sheltered English teachers and will disadvantage above average teachers, in accordance with the general principal that high performers should prefer the most accurate test possible whereas low performers should prefer a less reliable measure hoping to capitalize on chance. In addition to differential reliability, differential curricular alignment poses a potentially more serious measurement difficulty. Reading achievement is not unidimensional. English language arts (or English language development) objectives may be qualitatively different in sheltered English versus GATE classrooms, and it is unlikely that a single test would be equally well aligned with the curricula of these different classes. Teachers whose instruction is better aligned with the posttest will be relatively advantaged, because the test will be more sensitive to their students' gains.

To summarize, this simple case has highlighted four problems with teacher comparisons based on SGPs, even if a perfectly accurate pretest measure is assumed: (a) that students with the same pretest may differ in aptitude (and so may be expected to grow at different rates); (b) that students within a classroom will influence one another's growth (and so the presence of a few disruptive or high-needs students may distort the interpretation of the median SGP as a measure of teacher effectiveness); (c) that the posttest may measure more accurately for average classrooms than for those at the extremes; and (d) that the posttest may be differentially aligned with the instructional objectives appropriate to the needs of students in different classrooms.

The Second Scenario

Consider a system in which elementary schools are compared based on mean fourth grade reading achievement test scores, adjusted for third grade reading achievement. As before, assume that ideal, perfectly accurate third grade test scores are available for each fourth grade student, linked at the individual student level to fourth grade achievement test scores. As before, these data are used to calculate SGPs, and the median SGP is used to index the performance of each school. These school performance indicators should be superior to unadjusted mean fourth grade test scores, because the unadjusted mean would tend to be higher for schools serving more advantaged populations regardless of the quality of instruction.[3] Nonetheless, the median SGP is less than ideal.

Consider two schools, one an academic magnet school drawing students from a large attendance area and the other a traditional neighborhood school. Even after controlling for pretest score, pupils in the magnet school are likely to have parents who are more actively involved in their children's schooling and who provide greater out-of-school supports for learning as compared to pupils in the neighborhood

[3] If the third grade and fourth grade achievement tests are vertically scaled, then another possible school performance index would be the average gain score. Comparing schools based on average gain scores entails an assumption that the expected fourth grade gain should be the same number of points regardless of the third grade starting point. That assumption is not entailed when SGPs are used.

school. Thus, if the two schools are equally effective, pupils in the magnet school would be expected to have higher fourth grade reading scores even after controlling for third grade reading achievement.

As another instance, consider two schools with the same distribution of pretest scores but with markedly different student attendance and/or student mobility. Even after controlling for pretest scores, the school with better attendance or a more stable student population will have an advantage. The considerations cited in connection with the first scenario would also apply to this second scenario, as well. However, moving to the more macro level of school means instead of classroom means, individual differences in student aptitude would tend to diminish in relative importance, and so differences in out-of-school factors would loom larger by comparison.

The Third Scenario

Let us return to first case of a teacher accountability system, but extend that example to include teachers across multiple elementary schools. Consider two schools, A and B. In School A, the principal is more experienced, teacher morale is high, there is a clear and consistent school-wide focus on academics, and an effective information system keeps teachers apprised of student needs and progress. Classroom time is protected from outside interruptions. Students' class assignments are settled within the first week of the school year. Textbooks and other instructional materials are up to date and available in sufficient quantities to meet instructional needs. In School B, none of these conditions obtains. This contrast might be further elaborated by placing Schools A and B in different districts, with School A enjoying more effective district leadership and support in addition to its other advantages. The question then arises as to whether it is fair to compare teachers working under more versus less favorable conditions with no further statistical adjustments beyond controlling for prior year test scores.

To recapitulate, the first scenario of teachers in GATE versus regular versus sheltered English classes illustrates why control for prior achievement may be insufficient due to variation in student aptitudes, broadly conceived. The second scenario of magnet versus neighborhood schools or of schools with different attendance or mobility rates illustrates why control for prior achievement may be insufficient due to variation in out-of-school supports for learning. The third scenario of teachers working under more versus less favorable conditions illustrates why control for prior achievement may be insufficient due to variation in conditions generally regarded as under the control of the educational system but nonetheless outside the control of the individual teachers being evaluated. In each of these three scenarios, ideal definitions of *students similarly situated with respect to external factors* would incorporate additional variables beyond prior student achievement. The guiding principle in each case would be to define groups of similarly situated students in a manner that held constant precisely those achievement influences beyond the purview of the actors or institutions being compared or evaluated. Presumably, the remaining variation in end point achievement (beyond sampling and measurement error) would then be attributable to factors under the control of those actors or institutions. It must be emphasized that no bright line divides these two categories of achievement influences within versus beyond schools' or teachers' control. There is ample room for debate as to whether educators should or should not be held accountable for specific factors like student attendance, parent involvement, or student peer culture.

Recommendations for Measuring Student Growth

In practice, of course, reliable and valid measures of all these additional factors will rarely if ever be available. Even imperfect student growth measures may represent a dramatic improvement over unadjusted posttest scores. The point of the preceding thought exercise is not to conjure up the perfect as the enemy of the good, but rather to temper expectations as to what growth measures can accomplish. If students were randomly assigned to schools, teachers were randomly assigned to schools, and students were randomly assigned to teachers within schools, then most of these confounding factors would be expected to average out and strong causal inferences concerning teacher and school effects would be justified. In the real world, adjustments for prior achievement are helpful, but far from sufficient to assure fair and unbiased comparisons of teachers or schools.

Student growth measures should be designed for each specific application to provide the best possible control for influences beyond the purview of the schools or teachers being compared, subject to constraints of feasibility and data availability. Whenever possible, multiple prior observations should be used for each student (i.e., data from two or even three years earlier, in addition to prior year test scores). Growth measures over a longer time period (e.g., a full academic year) should be more stable than measures over a shorter time span. In general, results should be most accurate if prior year and end point scores are obtained at the same point in the school year. (For example, spring-to-spring comparisons would be preferable to fall-to-spring comparisons.) Vertically scaled measures are attractive in principle, but the measurement properties of vertical scales should be carefully examined, and to the extent possible, the strong assumptions entailed in constructing vertical scales should be tested. Scale invariant summaries, such as SGPs, may be used as the preferred achievement index or as a check on simpler gain scores or residualized gains. Median SGPs should be highly correlated with scale-dependent measures. Unless sample sizes are very large, however, gains or residualized gains may be preferable because they should yield estimates with smaller standard errors, especially if multiple prior years are included in the model.

Individual student growth measures should be continuous, not categorical, variables. Even if some need exists to set cut points and implement decision rules based on individual rate of growth, the growth measure should first be formulated as continuous. A given numerical value of the growth measure should have the same interpretation for all students. That means, for example, that if growth measures are to be aggregated across students at different grade levels, then measures expressed in the raw-score or scale-score units of different grade-level tests would not be appropriate unless the scale score were designed to provide common units across grade levels. Scale invariant measures like SGPs might be used, or grade-specific scale scores might be transformed to minimize distortions when scores are aggregated across grades.

Depending on the application, student characteristics in addition to prior test scores might be included. Parent education level, for example, might account for significant additional variance in end point test scores even after controlling for prior year scores. If, however, children of less educated parents on average receive poorer quality schooling, then inclusion of parent education in the growth outcome equation will adjust away a portion of the school effectiveness differences the growth measure was

intended to capture. Moreover, including this or other demographic variables may be misconstrued as setting different expectations for the achievement of different groups of learners. If two or even three prior years of data can be included, then reliance on other demographic variables is probably best avoided.

Defining and Comparing Growth for Groups of Students

Once individual growth measures have been constructed, the next step is to summarize those individual-level measures to define indexes at the level of classrooms or schools. This problem is not trivial, but it need not be very complicated. Problems arise in the construction of group performance measures because well-intentioned efforts to reflect educational policy priorities can result in the creation of measures with poor statistical properties. Percent-above-cut measures like *percent proficient* are a case in point. These indexes summarize group performance using the proportion of students with scores above some fixed cut score. They are what Holland (2002) referred to as *vertical gap measures*, and they often conceal more than they reveal about differences in group performance. Comparisons over time or among groups based on different cut points often yield wildly different results (Ho, 2008). In the context of growth measures, a similarly poor group-level index might be constructed by defining the minimum growth expected for each student and then estimating the proportion in a school that was meeting that expectation. Unfortunately, growth models implemented (under waivers) to meet the requirements of the No Child Left Behind (NCLB) Act may follow just this contorted logic, estimating the proportion of students on track to reach proficient by some future date. Percent-above-cut measures may be entailed whenever laws or regulations require the classification of students into discrete categories. Nonetheless, the replacement of continuous student-level scores with categorical classifications should be avoided whenever possible.

Simpler group-level summaries with better statistical properties are readily available. In the preceding examples using SGPs, median SGP was proposed, because SGPs, as with any percentile scale, have a rectangular, not a normal distribution. Instead of taking the median SGP, the SGPs might be transformed to stanines, or normal curve equivalents (NCEs), which could then be averaged. Gains or residualized gains can simply be averaged, as well, provided that the student-level index is expressed on a meaningful common scale for all students.

Group-level indexes created for accountability purposes often weight individual students differentially so as to create incentives toward some particular resource allocation within schools. This differential weighting might be accomplished by either of two methods. The most straightforward method would be to take a weighted average of the individual-level growth measures, with greater weight assigned to students for whom greater educational efforts were especially encouraged. Suppose, for example, that the goal were to encourage greater attention to students with lower pretest scores, so as to reduce the variation in achievement outcomes. Weights might be assigned based on quartiles of the pretest score distribution, say 2.00 for the lowest quartile and 1.00 for the remaining three quartiles. If a school had 20 students in the lowest quartile and 80 in the three highest quartiles, the school's average would be calculated as $(\sum w_i\, g_i)/\sum w_i$ where w_i was equal to 2 for each of the 20 bottom-quartile students and

equal to 1 for each of the remaining 80 students.[4] More complex examples might employ student weights based on multiple factors.

A second, less obvious method is in fact more widely used. This second approach is to employ a nonlinear scale for the individual-level index. That way, the overall index is more strongly influenced by growth in some regions of the achievement scale than in other regions. Several examples may be offered in the context of indexes based on status measures, but the logic in the case of growth measures would be exactly the same. The original score scale used in the Kentucky Instructional Results Information System (KIRIS) in the early 1990s is one example. Students were categorized on the basis of their test scores into four levels (novice, apprentice, proficient, and distinguished) weighted 0, 40, 100, and 140, respectively. Thus, schools were encouraged to work especially hard with apprentice students who were approaching proficient, because the greatest credit toward a higher school-level average score could be obtained by moving students past the proficient threshold than other thresholds. A more recent example is the use of *progressive scoring weights* in California's Academic Performance Index, which gives schools more credit for advancing students past the lower rungs of the achievement level scale (e.g., a 300-point difference moving from far below basic to below basic, but only a 125-point difference moving from proficient to advanced). A negative example is provided by the accountability provisions of NCLB. Even though NCLB requires that states define several performance levels, only the proficient level figures prominently in accountability calculations. Mathematically, the percent proficient metric may be described as the average of a transformed student-level score equal to 0 for below-proficient students and 1 for proficient students. This extreme nonlinear transformation creates an unintended incentive for schools to allocate resources toward *bubble kids* who are below but not too far below the proficiency threshold.

This second approach to differential weighting may be more politically palatable than the first, more transparent approach of explicit weighting. The explicit weights make it clear that some students count more than others in the group performance calculation. Nonetheless, a weighted average will have better statistical properties than a measure based on unequally weighted proportions of students in categories defined by cut scores. Particularly in the context of growth measures, any deliberate use of nonlinear scales may have complex consequences that are difficult to analyze or predict.

Another well intentioned but problematical mechanism for building policy priorities into group accountability indexes is the use of conjunctive decision rules. Under these rules some target must be met for each of two or more distinct subgroups within a school. In the context of status measures, the NCLB requirement that annual measureable objectives (AMOs) be met for the school as a whole and also for each numerically significant subgroup offers a familiar example. The same logic would apply if separate subgroup targets were established for growth measures. The goal of separately tracking traditionally underserved student groups is laudable, but a strong statistical bias is created when a system is designed such that sampling error or measurement error affecting any one of several small

[4] Score quartiles would be defined relative to the pretest score distribution pooling over all schools, so there need not be equal numbers of students in each quartile within any given school.

student groups can result in failure for the entire school. Separate reporting of results by subgroup is certainly sensible, but a single, simple group performance index would be preferable to a complex index with a conjunctive rule requiring that each of several distinct criteria must all be satisfied.[5]

Categories like basic or proficient are created in an effort to make test score results more interpretable, to say how good is good enough. They are then used to implement automatic decision rules specifying that one or another action be taken on the basis of achievement results. These same applications are likely with student growth measure summaries. There will be a demand for labels and categories as aids to interpretation and to implement policies. Judgmental standard setting methods have serious deficiencies even for characterizing the adequacy of achievement test performance at one point in time. Judgmental methods for characterizing the adequacy of growth should be avoided entirely. Ambitious but attainable targets should be defined empirically, based on the observed school-level distribution of the growth index, ideally over several years.

Regardless of whether continuous or categorical group-level indexes are used, it should be borne in mind that creating such an index cannot be value-neutral. Any possible index will represent a choice among alternatives embodying a particular set of assumptions and creating a particular set of incentives for educational practice. The individual-level index reflects values and beliefs about which achievement influences should be ignored and which should be accounted for when teachers or schools are compared. The group-level index reflects values and beliefs about which students' achievement gains should count most in evaluating a teacher's or school's effectiveness. Even the obvious default of equal weighting represents a choice among alternatives.

Using Group Performance Measures for Productivity Analysis

The previous two sections discussed the creation of individual-level growth measures and the use of such measures to construct group-level indexes. This third section considers how such group indexes might be used for productivity analysis. Applications of growth models for productivity analysis may be roughly divided into those comparing treatments (instructional programs, methods, or curricula, for example) and those comparing units (teachers or classrooms, schools, or districts, for example). With comparisons of treatments, the goal is to study a program or policy implemented across multiple sites. With comparisons of units, the goal is to compare particular sites to each other. Interpretive arguments differ somewhat for these two cases, but much of what has been learned over the years about social program and policy evaluation using nonexperimental data (the treatments comparison case) can be applied to evaluations of teachers and schools (the units comparison case).

[5] It should be noted that the multiple subgroup targets under NCLB also function as another indirect mechanism for differential student weighting. One student might be a member of the socioeconomically disadvantaged, English learner, and students with disabilities subgroups in addition to being a member of a group defined by one or another racial/ethnic category. The more groups a student is part of, the greater the impact of that student's performance in the school's adequate yearly progress (AYP) determination.

In establishing teacher or school evaluation criteria, unavoidable tension exists between holding all students' achievements to a common expectation versus recognizing that students differ systematically from classroom to classroom and from school to school with respect to aptitude, out-of-school supports for learning, and both prior and present schooling conditions beyond the control of the teachers or schools being compared. In the 1970s, annual (status) achievement score distributions for large urban school districts were interpreted relative to large-city norms for standardized achievement tests, offering a more favorable picture of student achievement than that portrayed by national norms. This interpretation may have been fairer to the school districts because national norms took no account of their particular challenges. However, the use of local norms may also have fostered complacency with low achievement and helped to perpetuate a culture of low expectations. The use of local norms of any kind has become less common as the accountability focus has shifted from the district level to the state level, but even now, states enjoy great latitude in establishing their own proficient definitions, and comparability across states is limited. Common standards may continue the trend toward greater uniformity in achievement expectations, but the tension between differential outcome expectations fairer to teachers and schools versus common outcome expectations expressing uniform societal goals for all students will not go away. A uniform achievement standard may well be trivial for nearly all students in the highest scoring schools and unattainable for nearly all students in the lowest scoring schools. Schools serving the lowest achieving students may be held to a standard of rapid achievement growth, but achievement gaps may be almost unchanged. With growth models, this tension may be obscured when expectations for schools or teachers are framed in terms of achievement gains or *value added* (acknowledging different starting points) versus attainment (e.g., percent proficient), but it will nonetheless remain.

Using Growth Measures to Compare Instructional Treatments

Current policy interest in growth measures centers more on comparison of units than comparison of treatments, but this brief comment on treatment comparisons may serve as a reminder of what is known about the limitations of covariance adjustments to compensate for preexisting group differences. Growth measures may not be described in these terms, but at bottom, individual or group level growth indexes are nothing more than covariance-adjusted outcome measures.

A textbook discussion of educational program or policy evaluations might begin with the ideal case of a true experiment, featuring random assignment of units (e.g., classrooms or schools) to treatment and control conditions. In this ideal case, unadjusted posttest measures suffice for comparing the treatment and control, although pretest scores or other covariates may be introduced to increase statistical power and precision. In the real world, random assignment of units to treatments is often infeasible or even impossible, and so the textbook discussion will move quickly to various quasi-experimental designs, using more sophisticated statistical models in an effort to control or adjust for any systematic differences between the groups of units included in one condition versus another. It is here that growth measures may be of particular value to help disentangle preexisting group differences from treatment effects. The problems of quasi-experimental design have been thoroughly studied for decades, and a large literature cautions against causal inferences from nonexperimental data unless heroic, usually

untestable assumptions are met. Growth measures may be used in such studies, but all the complexities of covariance adjustments for preexisting group differences will remain.

Many statistical problems identified in the context of quasi-experimental designs for program evaluation have close parallels in the use of growth measures for comparison of teachers or schools. As just explained, covariance adjustments generally cannot be relied upon to account for prior group differences in nonexperimental studies. Similarly, it has been emphasized in this paper that students with the same pattern of prior achievement test scores may not be equivalent for purposes of teacher or school evaluation. As another example, in curriculum evaluations, alternative achievement measures may be differentially aligned with one curriculum or another, in which case the rank ordering of curricular treatments may differ depending on which outcome measure is used (Walker & Schaffarzick, 1974). Similarly, as discussed earlier, an achievement test used for educational accountability may be differentially aligned with the instruction offered by different teachers or schools as they tailor instruction to meet the needs of students with different backgrounds and achievement levels.

Using Growth Measures to Compare Teachers or Schools

Comparisons of teachers or schools might be used to identify low performers for remediation or other special action, and to identify high performers for public recognition or for further study. Any such evaluation system must be carefully designed, because teachers and educational administrators may respond in both intended and unintended ways to the incentives the system will create. If teachers were simply compared on the basis of the end-of-year average student test scores, for example, they might compete for the strongest students within the school or withhold assistance from faculty colleagues. If teachers were persuaded that a student growth index equitably accounted for individual differences among students, these untoward consequences might be avoided, but that could be a very tough sell for teachers who know their students well and know which ones are progressing quickly and which slowly. Teachers will be aware of countless facts about the individual students in a school, facts that they will also know cannot be accounted for in the growth model—a troubled home, a chronic illness, an autistic sibling, or perhaps an extraordinary talent for mathematics or a special passion for reading. Comparing schools instead of teachers using a student achievement growth index would be much less problematical, because under such a system, all teachers within a school would have a common interest in maximizing the achievement of all students.[6]

By way of illustration, suppose that a student achievement growth index is created, and the average growth index is calculated for each elementary school in a state. At least three different *theories of action* might be set forth as to how such a school-level index could be used as a stand-alone tool to improve education. It will be argued, however, that more powerful school improvement models can be created if the growth index is incorporated into a more comprehensive educational indicator system.

[6] As noted later in this paper, limiting formal comparisons to the school level might not preclude the use of teacher-level data to apprise principals of persistent patterns indicating a need for coaching or some other remedy.

Before turning to indicator systems, three improvement models using just the school growth measure are briefly described: (a) accountability per se, (b) targeted resource allocation, and (c) studying outliers.

Accountability Per Se

Most current standards-based school accountability systems, both those using status measures and those using growth measures, are based on the notion that simply quantifying school performance and attaching rewards or sanctions to high versus low performance will lead to educational improvement. Several mechanisms for this hoped-for effect may be posited. One implicit theory of action simply holds that accountability will encourage teachers and principals to work harder.[7] Closely related to the idea of getting educators to work harder is the idea that public scrutiny can bring about school improvement. In the 1990s, many state systems relied on public reporting of school means to encourage parents to press for improvements in their local schools, and parent notification requirements are also included in NCLB. Accountability per se might also be expected to improve performance by encouraging better alignment of curriculum and instruction to the content standards embodied in accountability tests.

As thoroughly discussed by Koretz (2008), however, comparison of score improvements on high-stakes tests versus concurrently administered, low-stakes audit tests have repeatedly shown that the achievement gains commonly seen on high-stakes tests may be inflated and offer an overly optimistic picture of improvement in educational outcomes. Another justification for accountability per se is derived from economic principles. Various school choice plans have been formulated to challenge the perceived monopoly of public education and enable the working of market forces to encourage educational innovation and improvement. A market can only function if consumers have good information about the relative quality of competitors' products. Within this framework, accountability per se may bring about improvement by enabling parents to make informed choices concerning school quality. (Presumably, all parents should elect to send their children to whatever school has the highest test scores. Schools with persistently low test scores will either improve or else be forced to close because their students will go elsewhere.)

Targeted Resource Allocation

Moving beyond accountability per se, a school-level achievement or achievement growth index might be factored into decisions about resource allocations among schools. One approach has been to reward high achievement. Another has been to offer additional resources to schools found most in need of improvement. Both approaches are problematic. Rewards for high achievement may work against closing achievement gaps, as the rich get richer. Rewards for low achievement may create a paradoxical incentive toward poorer test performance. Most significantly, it may be difficult to assure that money allocated in this manner is well spent. Direct services to students are expensive, a recurrent cost that brings no longer term benefit for the education system. Professional development for teachers might

[7] If tests with stakes for individual students (e.g., high school exit examinations or tests required for grade advancement under the banner of "ending social promotion") are included, then incentives are created for students to work harder, as well.

improve performance, but research on the effectiveness of in-service professional development has found mixed results. Bonus programs may be popular in good economic times but are often cancelled when money becomes tight. The fundamental difficulty here is that school performance measures may signal that something is right or something is wrong, but they offer little insight into the problems or practices that might explain test score patterns.

Studying Outliers

An early model of school effectiveness research used regression analyses to predict average student achievement, identified schools with large positive residuals (i.e., schools beating the odds with observed performance exceeding expectations), and then followed up with case studies of the schools identified, seeking the keys to their success. In studies where prior test scores as well as demographic variables were included as predictors of current year test scores, the residuals used to compare schools were, in essence, growth outcome measures, although they may have been based on school-level data rather than student-level data. This research approach has fallen out of favor because it offers no statistical warrant for inferences as to the causes of unusual school success. Schools differ in countless ways, and the particular practices that the principal or other informants sincerely believe account for a school's exceptional success may also be found in other schools with poorer achievement outcomes. In short, identifying high performing schools and then, after the fact, looking for explanations for their high performance is a poor way to do research. Moreover, residualized gains are often highly unstable. Different schools are likely to show up as outliers in different years or using data from different grade levels or subject areas (Ma, 2001; Mandeville, 1988; Mandeville & Anderson, 1987).

Using School Growth Measures in Educational Indicator Systems

As shown by the foregoing discussion, student test scores have some utility in stand-alone applications. Arguably, schooling practices and outcomes have improved as a result of accountability per se as well as targeted resource allocation, and these elements are likely to appear in future educational policy initiatives. Both can be done better using growth measures to index schooling outcomes. Growth measures can also be used in educational evaluations (e.g., comparisons of curricula or teaching methods) and quantitative research studies (e.g., examinations of large-scale educational policies), reducing bias and improving precision to clarify findings and interpretations.

All of these improvements upon the status quo are worthwhile. However, significant, systemic improvement in our educational system will require not only improved measurement of schooling outcomes, but also a deep analysis of the relationships between outcomes and a range of contextualized schooling policies and practices. This kind of analysis is essential to identifying the causal relationships that must be understood to inform wise policies. It will require the assembly of collections of quantitative measures describing practices as well as outcomes, deliberately designed with specific goals and linked at the student, teacher, and/or school levels. Such a collection of measures is referred to

here as an *educational indicator system*.[8] A better understanding of those factors that teachers, principals, and education policymakers can manipulate is clearly essential for dramatic school improvement.

An educational indicator system cannot be comprehensive. It may be narrow or broad, but in every case it is systematically designed for a specific purpose such as diagnosis, monitoring, evaluation, or explanation, with logical connections hypothesized among measures of school characteristics, schooling policies and practices, and growth outcomes. Some indicator systems will last the duration of a research study; others may be intended to last indefinitely, for ongoing monitoring and management. This paper will conclude with just one final example of an indicator system for ongoing monitoring of educational productivity at the school level, implemented in a (hypothetical) large urban school district or a consortium of districts with at least 200 elementary schools. The school level is chosen for this example to avoid the problems of competition among teachers within a school. Although the primary purpose is ongoing monitoring, the indicator system would also provide information over time that could be used to support generalizable conclusions about best practices.

One primary outcome measure in this system would be a school-level growth index based on annually administered standards-based achievement tests aligned to the state's curriculum frameworks (or to common core state standards) and linked over time at the individual student level. However, the indicator system would also include two additional categories of student outcome variables as well as additional indicators of student characteristics and of schooling practices. With regard to schooling outcomes measures, the growth index would be complemented by cross-sectional (status) measures of student achievement, derived from the same annual achievement test data. In addition, to compensate for widely recognized limitations of standardized test scores alone, additional outcome measures would be included, based on students' work in the classroom. These would be evaluations of extended projects that, even at the upper elementary school level, would require students to find, organize, and evaluate information; to analyze and synthesize data; to plan and carry out investigations; to solve unfamiliar, open-ended problems; and to write reports and give oral presentations. Evidence from these activities might be assembled into student portfolios, but only if teachers found portfolios to be pedagogically useful. Detailed specifications would be provided to schools in the district, specifying the general features and the cognitive demands of the activities required, but within each school, teachers would have considerable latitude in designing the assessments themselves. Several years would be devoted to phasing in these activities in one subject area at a time before the scores were actually used as part of the indicator system. Teachers would score the tasks locally within each school, but committees with representation from multiple schools would afford some external review of tasks and scoring criteria. The cross-sectional standardized measure that was common across all schools would be used to place within-school extended task scores on a common scale via *statistical moderation* (Linn, 1993; Mislevy, 1992).

[8] It is somewhat ironic that the argument developed in this paper, beginning with the analysis of individual growth measures in the first section, focused on achievement influences *outside* the school's control prior to any consideration of achievement influences *within* the school's control.

Growth outcome and status measures derived from standards-based tests and classroom-based assessments derived from extended tasks would offer comprehensive school-level information on student learning outcomes. Two more categories of variables would complete the indicator system. First, in order to further contextualize the growth outcome measures, additional student information variables would be included. Second, variables characterizing schooling policies and practices would be included, so that variations in practice could be related to patterns of schooling outcomes.

Growth outcome measures use prior year test scores as proxies for all the test score influences that are outside the school's control, but as discussed earlier, even students with identical prior year test score patterns may differ in their readiness to profit from further instruction. Student age in months might be used as an additional explanatory variable. Other things being equal, between a younger student and an older student with the same test scores, the younger student has grown more rapidly and may be expected to continue to do so. Student information variables would also characterize students' prior instructional histories. Year and month of matriculation at the current school as well as year and month of first enrollment in the school district, prior year attendance, English language status, and flags for disabilities are all obvious measures that should be readily available. Instructional history might also include prior year scores on school-level extended tasks. Racial/ethnic classifications, gender, and other group-level characteristics would not be used.

Schooling policy and practice variables would be the most extensive and the most complex category included in the indicator system. Teacher IDs would be linked to each student record, so that principals could be informed of teachers whose students were consistently low-performing or persistently high-performing year after year, but strong safeguards would assure that those data were used solely for coaching or remediation. Finally, to the extent possible, whenever the district undertook a curricular innovation, in-service activity, or initiation of a new policy (e.g., parent participation contracts), the initial implementation would be for a random subset of schools (typically around 50%). If the new policy or practice were voluntary, then schools would be asked to volunteer, and a fraction of those volunteering would be randomly selected for the initial implementation. For each such innovation, indicators for volunteering as well as selection would be added to the indicator system. At the time the policy or practice was initiated, a date would be specified when it was to be evaluated. In this way, the indicator system could be used for ongoing, routine evaluation of policies and practices to learn what was actually working or not working in that particular school district. Innovations shown to be effective would then be adopted more broadly. Those not shown to be effective would be terminated.

Summary and Conclusions

Growth outcome measures are a useful complement to status outcome measures of learning. Many current and historical uses of test data for educational improvement, including program evaluation, data-driven resource allocation, and what was here termed accountability per se, may be improved by using growth outcomes in place of cross-sectional outcome data. Growth outcome measures hold still greater promise as part of educational indicator systems designed to investigate the effectiveness of programs and practices. There is no one right way to build such a system, and an indicator system built for one purpose is likely to be suboptimal for others. A brief sketch of one such system, for elementary

schools in a large district, included school-level growth and status outcomes derived from a statewide standards-based test, local classroom-based measures of valued learning outcomes not captured by the external standards-based test, indicators of student aptitude (broadly conceived to include many kinds of indicators of readiness to profit from further instruction), and a flexible and growing collection of indicator variables corresponding to programs, policies, and practices adopted by the school district.

Effective use of information from any such system will require getting the relevant information to the right people at the right time to inform actual decisions. This is more than a measurement problem. It requires attention to the larger, technologically supported social system linking students and teachers to administrators and policymakers at the school and district levels and beyond, and it requires sustained attention to the problem of creating and maintaining a district-wide culture of data-driven decision making.

Understandably, policymakers often defer to technical experts, especially with regard to complex statistical matters. This response is perfectly appropriate. What is not appropriate is for policymakers to expect (or demand) technical answers to value-laden policy questions. In matters of test use and interpretation, the proper role of technical experts is to clarify choices and associated trade offs and to facilitate wise decisions to maximize benefits and minimize unintended negative consequences. But the valid use and interpretation of student achievement data for educational monitoring and improvement will require collaboration among many actors within the system. Growth measures may be useful tools in support of these goals, but they are far from a ready-made solution.

References

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*, 351–360.

Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, *27*(1), 3–17.

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83–102.

Ma, X. (2001). Stability of school academic performance across subject areas. *Journal of Educational Measurement*, *38*(1), 1–18.

Mandeville, G. K. (1988). School effectiveness indices revisited: Cross-year stability. *Journal of Educational Measurement*, *25*(4), 349–356.

Mandeville, G. K., & Anderson, L. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of Educational Measurement*, *24*(3), 203–216.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods and prospects*. Princeton, NJ: ETS. (ERIC Document Reproduction Service No. ED353302)

Walker, D. F., & Schaffarzick, J. (1974). Comparing curricula. *Review of Educational Research*, *44*(1), 83–111.