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

Measuring Student Growth With Large-Scale Assessments in an Education Accountability System

Wendy M. Yen

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.

Measuring Student Growth With Large-Scale Assessments in an Education Accountability System

Wendy M. Yen

Educational Testing Service, Princeton, New Jersey

This paper is based on a reaction by Wendy M. Yen to presentations by Damian Betebenner and Robert Linn at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download copies of the papers presented at the seminar at http://www.k12center.org/publications.html.

I am very pleased to have been invited to discuss the presentation by Damian Betebenner and Robert Linn, *Growth in student achievement: Issues of measurement, longitudinal analyses, and accountability.* I will first comment on their presentation in general and then focus on some ideas that I believe are particularly relevant to the U. S. Department of Education's Race to the Top agenda and the use of growth measures in large-scale, accountability settings.

The Betebenner and Linn presentation is a tour de force analysis of the quantitative and policy aspects of measuring growth in educational achievement. They separate out three important, overlapping views for considering growth: psychometrics, applied statistics, and accountability. By considering these three perspectives separately, their presentation brings new clarity to the discussion of growth measures in educational accountability. Professor Linn's uniformly sound judgment based on his decades of thoughtful work in educational assessment is manifest in the wise advice that permeated this presentation. And I have been enormously impressed with Damian Betebenner's growth percentile work as implemented in Colorado. Not only has he made sound, innovative technical decisions, the collaboration with Colorado educators has produced outstanding communication with users.

In writing my comments, I have kept in mind the context for this workshop, which is to catalyze higher quality assessments within Race to the Top applications. Toward this end, my discussion will cover some general concepts about assessing growth in large-scale assessment and then turn to some observations about educational accountability. I end with a short list of Race to the Top recommendations.

Growth

The development of student abilities in reading, mathematics, and other academic areas is a complex, multidimensional process. If one is concerned with helping students learn in the classroom, it is the "now" and the "what" of student performance that are important—it is the real-time qualitative observation and interpretation of academic content and evidence of student knowledge that is useful. The quantitative evaluation of *how much*, especially if that *how much* is spread out over a year, does little to help individual students learn. Later presentations focus on formative assessments, which

directly address these classroom issues. For now, I am going to focus on high-stakes summative accountability testing and its measurement of growth.

The No Child Left Behind (NCLB) growth pilot has resulted in a wide variety of innovative proposals for measuring or projecting growth relative to adequate yearly progress goals (http://www.ed.gov/admins/lead/account/growthmodel/index.html). Some states use a vertical scale and some do not. Some count changes in students' proficiency level classifications and some apply numerical values to such changes. There are states that use some type of regression for relating student performance between grades; in some cases the regression is focused on projecting whether a student is likely to be proficient in a certain number of years and in other cases growth norms are developed. There is not time here to describe and to cover the pros and cons of all the different models; such a discussion is contained in my paper, *Growth models approved for the NCLB growth model pilot* (Yen, 2009). I will instead skip to my major conclusions and advice about incorporating growth models into Race to the Top accountability models.

As Betebenner and Linn discussed, different questions can be answered from different types of data and analyses. Examinations of growth can be a valuable addition to status models in reviewing progress. While growth models should be encouraged, they should not be mandated to the exclusion of status models. Neither should one particular growth model be mandated. Different models can be suitable for answering different questions, and mandating one model would preclude the type of creativity that has arisen in the NCLB growth model pilot. Again, as Betebenner and Linn stated, for any type of analysis, be it based on status or growth, the measurement properties of the procedure should be commensurate with the requirements of the analysis.

An important aspect of growth models is considering vertical scales. Following is a brief description of vertical scales from Yen (2009):

> One psychometric test property that can facilitate the measurement of student growth is a vertical scale. In a vertical scale, for each content area (e.g., reading or math), scale scores (typically three-digit numbers) are produced that run continuously from the lowest grade tested to the highest grade tested, with substantial overlap of the scale scores produced by adjacent grades. On the vertical scale, Proficient might be 350 in grade 3, 380 in grade 4, and 400 in grade 5, and so forth. The difference in a student's scale scores at adjacent grades is a measure of the amount of academic growth achieved by that student.
>
> With a vertical scale, an ideal goal is to have scale scores that have the same meaning if they are obtained from different test levels (e.g., a 400 *means the same thing* or represents equivalent knowledge or achievement whether it comes from the grade 4 test or the grade 5 test). Also, differences between scores are ideally comparable for gauging amounts of academic growth. For example, a student who grows from 350 to 370 (20 points) would be

> demonstrating the same amount of growth as a student who grows from 390 to 410 (20 points).
>
> To produce a vertical scale, it is assumed that the tests at adjacent grades have substantial overlap or articulation in content and that a single major dimension (e.g., overall mathematics achievement) explains most performance differences. If a state's standards, curriculum, and assessments are designed to have large distinct subareas of content that are not designed to be taught or learned hierarchically, then a vertical scale is not expected to produce sensible results…. Vertical scaling makes the implicit assumption that the same construct is being measured at the top and bottom of the scale. This assumption may be difficult to justify when the vertical scaling includes many grades. (Yen, 2009, pp. 3-4)

As someone who has been constructing vertical scales for more than 25 years (Yen, 2007), I would expect that if Common Core State Standards are developed with vertical articulation of content in mind, and these standards and assessments are connected to actual curricula and instruction, vertical scales with intended ideal properties can be created—but only approximately. No matter the assessment or the details of the scaling procedure, as grades get farther apart, and the tested content diverges and changes its dimensionality over grades, the entire concept of equivalence of scores and score units does not have a logical basis, and is, in fact, untestable. (For example, it doesn't make sense to give Grade 3 students a Grade 7 test to see if equivalent scores are produced.)

Vertical scaling can produce approximate linking of performance across grades—and this approximate linkage can be useful for low stakes scores (as it is in producing grade equivalents for norm-referenced tests) and for, say, visual displays of results (as is done so effectively with the growth data on the Colorado website). But the untestable assumptions that are inherent in a vertical scale that covers many grades make the results inappropriate for high-stakes usage that assumes that 10 units of growth for a low-scoring student is "equivalent" to 10 units of growth for a high-scoring student or that 10 units of growth for a third grade student can be precisely compared to 10 units of growth for a seventh grade student. The results of high-stakes summative assessments are scrutinized in minute detail and small differences can have big policy implications. Vertical scales, which are based on untestable assumptions, cannot be demonstrated to have the accuracy needed for wide-ranging applications in high-stakes accountability testing.

In thinking about vertical scales, the phrase *good enough for government work* came to mind. According to *Wiktionary*, this phrase, "Originated in World War II. When something was 'good enough for Government work' it meant it could pass the most rigorous of standards. Over the years [this phrase] took on an ironic meaning that is now the primary sense, referring to poorly executed work" ("Close Enough for Government Work," n.d., Etomology section, para. 1). I would like to be sure that the high-stakes educational accountability measures that we endorse are *good enough for government work* in the original, rigorous sense.

Henry Braun pointed out more than 20 years ago that comparisons of growth can be made most accurately when comparing outcomes for students who start at the same place. When students are starting at different places on the scale, differences in the scale units can cause distortions in conclusions (Braun, 1988). Longitudinal regression, in which student performance is tracked from one grade to the next, is an excellent way to measure and compare growth. Regressions permit the comparison of outcomes for students who start in the same place (that is, with the same score at the end of the previous grade). In that sense, they are merely descriptions of reality—what actually happened for students going from one grade to the next. Regressions do not require a vertical scale. However, a regression can be used with a vertical scale, and by taking starting places into account, regression avoids the pitfalls of the untestable assumptions of the vertical scale. Regressions come in different flavors—empirical or linear, scale–score-based or quantile-based, univariate or multivariate, and so forth. One can debate the pros and cons of the different variations, but—except for the most complex regression models—they all share the advantage of measuring growth with minimal assumptions.

As Betebenner and Linn pointed out, there are both absolute and relative aspects related to our understanding and interpreting student growth. In the absolute sense, we ask whether a student has reached the state-defined level of *proficient* performance. In the relative sense, we also care about comparing our student's growth with that of other students, which is, in essence, a normative interpretation. In 2005, we explored California educators' attitudes toward different growth measures and found that while they had many questions about student growth that were implicitly normative, they were also concerned that discussion of norms would draw attention away from the absolute goal of all students becoming proficient (Smith & Yen, 2006). The Betebenner growth percentiles approach in Colorado, which has focused so successfully on communication of results, has done an outstanding job of providing useful normative growth information without diminishing focus on the absolute goal of students becoming proficient.

Some Observations on Educational Accountability Systems

I am no expert in educational policy. I have no training in that area, nor do I do research in it. However, I have been involved in educational accountability testing for more than 30 years and have seen many systems put in place. I would make the following observations: (a) Everyone in education wants to help children learn, (b) opinions are strongly held about how to help children, (c) opinions differ about how to help children, and (d) there is conflict in educational policy. I have also observed that educational systems are very complex. They are reactive and dynamic; their social, economic, and political surroundings are constantly changing and being changed by education. While everyone is well-intentioned and we are all doing our best, all the implications of any of our actions or changes cannot be foreseen.

Within this complex system, what can accountability testing contribute? It can spur thought into the meaning and alignment of content standards, curricula, instruction, tests, and performance standards. And it can help define and increase focus on what we want our students to learn. Accountability testing can help us see what change is taking place and what change is not. For example, in California from 2003

to 2009 we saw steady, sustained improvement in the percentages of proficient students in English language arts and mathematics. We have also seen that the gaps in performance between African American and White students, and between Hispanic and White students, persist virtually unchanged. Without accountability testing, we would all be guessing about what change has happened and what change has not.

Historical accountability testing information can also provide a setting in which we can evaluate the difficulty of goals and consider the resources needed to reach them. For example, imagine that historically the largest sustained increase in percentages of proficient students for states across the nation has been in the range of 2 to 4% per year.[1] California's improvement has been in that range, yet the state's NCLB challenge was an improvement of 6% per year sustained for 11 years resulting in 100% of students becoming proficient. Such exemplary performance has never been observed for any state, which would indicate that extraordinary resources and actions would be needed in attempting to meet that goal.

Performance goals can be idealistic and developed with the best intentions, but if they are unrealistically high they are counter-productive. Betebenner and Linn clearly described the unintended negative consequences that can occur with high-stakes accountability systems: narrowed curriculum, gaming the system, emphasis on test preparation, and demoralization of those most involved in school improvement efforts. I believe these consequences are especially likely to occur when a system has unrealistically high goals.

Historical accountability testing information can be used to help us create stretch goals that *are* attainable. For example, we might look at the distribution of school results and identify those schools with the top 25% improvement. Not only could the amount of improvement seen by those top schools be used as a stretch goal for lower performing schools, the top schools' best practices could be shared. In my work with California educators, I have found them eager for such information.

Recommendations for a Revised Accountability System

1. Broaden the measures of achievement and success beyond one summative, high-stakes assessment.

   The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) stated (p. 167), "Standard 15.4 In program evaluation or policy studies, investigators should complement test results with information from other sources to generate defensible conclusions based on the interpretation of test results." In other words, consideration of multiple measures can contribute to the validity of conclusions drawn about educational achievement (Henderson-Montero, Julian, & Yen, 2003).

2. Develop goals for improvement that are challenging but attainable.

[1] Circa 2001, an informal survey of states found these results (Schwarz, Yen, & Schafer, 2001). I have not surveyed current state results.

Historical information about school and student performance is very important in the development of these goals.

3. Measure and equate accurately.

   Whatever accountability measures are put in place, they will be the subject of intense focus and minute scrutiny. Small changes in percentages of proficient students can have major consequences. If we are to measure change using educational assessments, it is critical that the assessments have sufficient measurement quality that they can be very accurately equated. Psychometricians know how changes in assessments can affect measurement accuracy and equating—changes such as modifications in content coverage or item formats or passage length, changes in the severity of raters for constructed response items, changes in the number of items or timing, changes in item context or item positions, changes in testing conditions or motivation, and so on. I believe that accurate equating is the most important service provided by psychometricians in large-scale assessment, and the requirements for equating cannot be taken lightly.

4. Encourage growth measures that do not rely on untestable assumptions and do support both absolute and relative (that is, normative) interpretations.

   Vertical scales can provide useful, approximate information about relationships among scores across a range of grades. However, vertical scales that cover many grades depend on largely untestable assumptions, and they do not produce results that have the precision needed to develop dependable conclusions in wide-ranging high-stakes applications. Longitudinal regressions, which take students' starting places into account, can provide precise growth measures for both absolute and relative interpretations and avoid the pitfalls of vertical scales.

5. Be cautious in attributing causation to results, be they status or growth results.

   As Betebenner and Linn (and many others, such as Braun [2005]) have pointed out, unless students are randomly assigned to teachers, there is no solid statistical basis for attributing students' growth to individual teachers' actions.

6. Emphasize communication of results to users.

   For testing results to be useful, they must be communicated in a way that educators understand. Verbal and graphical examples of appropriate use and interpretation are essential. The Colorado website (https://cdeapps.cde.state.co.us/growth_model_public/) has outstanding examples of multiple, effective ways of communicating growth results to constituents.

7. Innovate responsibly.

   I think we all tend to have the belief that if there is a problem or issue, and we are attentive and intelligent in our thinking, we can take an action or make a change that will improve things. In many cases that is true, but in some cases it is not. For example, in the past couple of years I attempted to keep an eagle eye on my retirement savings. I attended to a variety of reputable advice—and in fact I kept a folder full of relevant articles and clippings of interest. After careful deliberation, I took

thoughtful action based on some of this advice— and I could not help but notice that all my actions were not wholly successful. Recently when I went back to clean out that folder, I was struck by how many of the articles, which looked so sensible at the time, were wrong.

In educational accountability testing, I have seen many innovations where dedicated, intelligent educators have believed that this next innovation will be the one that changes everything for the better. Many people are affected by our actions in accountability testing. People in positions of power—those in U.S. and state departments of education and legislatures, chief state school officers, as well as school board members—rely on us as measurement professionals to give them honest advice and do our best to help them do the right thing. Many of those affected by accountability testing might be called *the little people*—the teachers and school administrators who have minimal say in designing accountability systems but who will be the ones carrying the ball and under intense pressure to look good on whatever measures are put in place. Given that we cannot anticipate all the effects of any of our actions, we owe it to all these people, and most importantly to students and society as a whole, not to be overly confident in our innovations or to oversell them.

As we explore the next generation of K-12 assessments, no doubt a variety of new, creative ideas will be advanced. This is all well and good. However, when psychometric innovations are implemented in high-stakes settings, it is incumbent upon us to demonstrate *before* the implementation that the measurement properties of the system, in particular the equivalence and comparability of scores, are sufficient for their intended use.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Braun, H. I. (1988). A new approach to avoiding problems of scale in interpreting trends in mental measurement data. *Journal of Educational Measurement, 25,* 171-191.

Braun, H. I (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: ETS.

Close enough for government work. (n.d.). *Wiktionary.* Retrieved from http://en.wiktionary.org/wiki/good_enough_for_government_work.

Henderson-Montero, D., Julian, M. W., & Yen, W. M. (2003). Multiple perspectives on multiple measures: An introduction. *Educational Measurement: Issues and Practice, 22*, 6.

Schwarz, R. D., Yen, W. M., & Schafer, W. D. (2001). The challenge and attainability of goals for adequate yearly progress. *Educational Measurement: Issues and Practice, 20,* 26-33.

Smith, R. L., & Yen, W. M. (2006). Models for evaluating grade-to-grade growth. In R. W. Lissitz (Ed.), *Longitudinal and value added modeling of student performance* (pp. 82-94). Maple Grove, MN: JAM Press.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J.Dorans, M. Pommerich, & P. W. Holland, (Eds.), *Linking and aligning scores and scales* (pp. 273-282). New York, NY: Springer.

Yen, W. M. (2009). Growth models approved for the NCLB growth model pilot. Unpublished manuscript.