



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

# Comments on Growth in Achievement

Mike Kane

*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.*

Copyright © 2010 by Educational Testing Service. All rights reserved. ETS is a registered trademark of Educational Testing Service (ETS).



## Comments on Growth in Achievement

Mike Kane

Educational Testing Service, Princeton, New Jersey

This paper is based on reactions to presentations by Damian Betebenner and Robert Linn and by James Pellegrino and to comments by Wendy M. Yen 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>.

The presentations by Betebenner and Linn and by Pellegrino and the comments by Yen were very extensive, very thorough, and masterful. I did not find anything with which I disagreed, and I found lots of places where I said, “Great!” I will simply mention a few points made in the presentations that I would like to re-emphasize, and make a few additional comments.

### Growth in What?

Both presentations talked about the kinds of growth in student learning we should seek. This is a core issue, if not the core issue. The problems with scaling and equating for measuring growth in student learning are serious, but we must start by defining the kinds of growth that we want to see. We have sometimes thought of growth as *knowing more* in a sense—knowing more facts, more skills, more theories, and so on—without necessarily focusing on the kinds of competence or expertise involved in being able to perform effectively in some area of activity or some content domain. There is some danger that, in developing psychometric models and measurement procedures, we will get caught up in technical problems, and thereby lose sight of the kinds of growth that we want to encourage. Technical problems tend to become more tractable as we simplify our conception of what we want to measure, and so there is a tendency to focus on narrowly defined domains. Presumably, we do not really want to produce a generation of savants who know a lot of scientific detail but do not have a good sense of what science is all about, what it means to do an experiment, and how scientific principles play out in refrigerators and light bulbs, and in how to bake a cake. We want students, who generally start as novices in academic domains, to develop competence and/or expertise in these domains.

### Models for Growth

I like the use of learning progressions as a basis for measuring growth because it seems possible that, by building measurements that go with learning progressions, we can tie the assessment results to important qualitative outcomes that parents, students, and teachers should be able to identify with. The things we want students to be able to master are not simple traits, and are not going to be represented by a simple unidimensional construct. They are more like cluster concepts, involving a set of related

skills, knowledge, and competencies that go together. Where we can differentiate different levels, the differentiation is not on a single simple dimension, but on a cluster of competencies. It seems that the learning progressions capture some of this kind of complexity and are particularly appropriate for assessments of growth, because they provide a rich conception of what we want to measure in evaluating growth. In particular, they emphasize that growth is not unidimensional.

In the display with the trees, presented by Jim Pellegrino, I would like to see different kinds of trees. As it was, a single kind of tree kept getting bigger, but in point of fact, in real woods, some trees grow very tall, very quickly, and some find niches between the larger trees. Depending on where they are, trees may grow to one side seeking the sunlight. We see very different patterns of growth in trees, and we see an equal or greater range of growth patterns in students in school. Some students in a math class may get to be very good at computations but not really understand what the computation is about, while others may find structural patterns fascinating. Future mathematicians are likely to come from the latter group.

The students who did not really understand the physics they were studying, even though they could solve the problems and use the equations, were not unusual. Richard Feynman gives a similar example of this phenomenon in one of his books; his students can derive complex conclusions about polarized light from the mathematical models, but do not recognize an example of polarization when it is in front of them.

I have not been optimistic about the utility of cognitive models of performance for assessment in most areas of the curriculum, at least in the short term. Most of the cognitive models of performance that I have seen are in areas that are almost algorithmic (e.g., some aspects of math and science such as solving problems with a given structure). Not only do we not have detailed cognitive models for U.S. History, I cannot imagine what such a model would look like. Cognitive models of performance look at how people do the kinds of things they do, which is somewhat different from the learning progressions. The learning progressions seem to be defined in terms of the structure of the content domain—they focus on how different constructs, methods, and assumptions fit together, vertically and hierarchically, rather than on what is going on in the student’s head. The learning progressions can be developed more easily than cognitive models and can have different levels of granularity, coarse or fine. In most areas, cognitive models will be most useful in providing general insights about learning that help us to develop realistic learning progressions, which can then be used to guide test development.

## **How Much Growth?**

The next issue after the growth-in-what question is how much growth do we have? This is a tricky question, because growth is generally not unidimensional, and a simple “how much” suggests a unidimensional quantitative scale. How do we measure growth, if achievement is multidimensional?

If we can lay out development along a single line, we can evaluate how far people have moved along the line, but these analogies with length measurement are somewhat misleading. In many areas of science, good measurement involving unidimensional scales did not occur until after the phenomena had been

studied for decades, if not centuries. Temperature scales may seem simple, but they are based on a lot of sophisticated theory and research.

In most areas, the learning progressions are going to be curriculum-bound, in the sense that the progression depends in large part on how we a) define the domain and b) sequence the topics in instruction. There may be a few areas where you could not learn the next thing without knowing the previous thing, but in most areas there is not going to be a single unidimensional set of topics (like beads on a string) that have to be learned in lockstep, so the actual progression is going to depend on the curriculum and the student. One student may be looking out the window when we are talking about the quantitative ways of describing chemical reactions, but really love working in the lab with test tubes, and another student may love the models and hate the lab. We need lots of people with different interests and skills in our society.

## Scaling

I like the quote from Tukey in the Betebinner and Linn presentation to the effect that it is better to be vaguely right than precisely wrong. I was especially taken by that quote, because I was reminded that my advisor in graduate school, Patrick Suppes, trained as a logician and philosopher of science, was fond of saying that it is better to be precisely wrong than vaguely right—the exact opposite of the sentiment expressed by Tukey. Many maxims are like that—they come in pairs that express opposing sentiments: A stitch in time saves nine, but do not cross the bridge till you come to it. Do not put all your eggs in one basket, but nothing ventured, nothing gained. We have two statements that express opposing approaches to life and we accept both as guiding principles. How do we get away with that?

To some extent, it is a matter of balance. Preparing for possible problems in the future is important, but it is also important to live in the present. It also depends on the context. Suppes was talking about logic and the foundations of science, about conceptual analysis. In that context, being precisely wrong is better than being vaguely right, because if you are precisely wrong, you are likely to be able to figure out that you are wrong, and perhaps figure out something about how you are wrong, and then you can do something about it. In particular, you can look for a new theory. It has been argued that being precisely wrong in this way is the main impetus for major progress in science (i.e., paradigm shifts). If you are vaguely right in science, you do not have any reason to change anything. In addition, philosophy and science take the long view: If we get it wrong this decade, but fix it next decade, that is fine.

On the other hand, in an applied context (e.g., in an educational setting with real teachers and students), the consequences of being precisely wrong can be serious. If we make a hash of fourth-grade instruction this year, and fix it next year, we cannot easily repair the damage to this year's fourth graders. So, in these settings, it would generally be better to be vaguely right than precisely wrong. In any case, we have to live with both admonitions at the same time.

In this vein, for some purposes, it is useful to pretend that our assessment results are unidimensional and that what we are measuring is unidimensional. In large-scale, standardized testing, unidimensionality assumptions can be both plausible (approximately right) and very useful for scaling and equating. For classroom assessment and, more generally, for diagnostic testing, unidimensionality gets in the way, and

is clearly wrong as a detailed description of reality. As Box was quoted as saying, all models are wrong, but some tend to be useful. So, in defining our conception of the kind of achievement in which we want to see growth, we may need to make our assessment unidimensional enough that we can scale and equate the scores, but multidimensional enough that we care about growth.

Very detailed descriptions of performance may be very useful in the classroom or the laboratory, but they are not so useful in describing group-level performance or in many decision contexts. It comes down to a question of grain size, and my sense is that the learning progressions that were discussed in the two presentations are at about the right grain size for measuring academic growth over extended periods (months or years, rather than day to day). Furthermore, the grain size can be adjusted if necessary.

Assuming that we have a learning progression at more or less the right grain size, we may want to define our scale units at about the same grain size as the learning progression. That is, if we have a learning progression with 12 general levels, we could accomplish a lot by focusing on a scale with 12 points, 1–12 (or 10–120, if we want to preserve the opportunity to consider intermediate levels without decimals). One obvious (and major) advantage in this approach is that it emphasizes the connection to the learning progression and encourages interpretation in terms of the learning progression.

In any case, our measurements are generally not very precise, and therefore do not support fine distinctions. Our psychometric models generally assume some underlying continuous scale (e.g., theta or true scores), but at fine grain levels, many of the assumptions in these models (particularly unidimensionality) break down. So we give up little if we focus our attention on the scale points associated with the major levels in the learning progression. We may want to have finer-grained scales in the background for technical reasons, but reporting results, in terms of a dozen levels in a learning progression, is probably as meaningful as we get in most cases.

In the psychometric community we value precision and getting more and more accurate descriptions of whatever we are measuring even though we do not know exactly what the construct is. When I think about improving reliability from 0.89 to 0.90, I am often reminded of the old joke about the airline pilot who comes on the public address system and says, “We have good news and bad news. We have a strong tailwind and are making excellent time, but our compass is out, so we are lost.” We often have a lot of precision but we do not know exactly what we are being precise about.

Furthermore, as noted above, our measurements are not as precise as we pretend that they are. When we assess reliability, we typically only look at one or two sources of error, and we may find that we have a high reliability. In the physical sciences, there has been more attention to measuring every conceivable source of error. When you do that, you find that your estimated precision is lower than it is when you just focus on one or two sources of variability (e.g., internal consistency or rater agreement).

Again, for some purposes (e.g., equating, scaling), we may want to have a finer-grained score scale in the background. If we do not do that, we can lose a lot of precision (because of rounding) without getting anything in exchange. I am not against precision; I just do not want to pay too much for it.

## Technical Criteria

As discussed above, a good prima facie case could be made for the validity of assessments based on learning progressions as measures of status on the progression and as measures of progress on the progression. At least at a general level, assessments that are built to reflect the learning progressions, and are scaled to reflect the progressions, provide a basis for drawing conclusions about a student's level of achievement in the progression. If a student can handle most of the tasks at level 5 and few of those at level 6, a conclusion that the student is functioning at level 5 seems reasonable (assuming that there is no evidence that contradicts this conclusion). Furthermore, to the extent that levels 4 and 5 in the learning progression are well defined, and the score scale is well linked to the learning progression, we have a good idea of what it means to go from level 4 to level 5. I recognize that I am not saying anything new here; this advantage in interpretability is clearly one of the main reasons for developing learning progressions that are linked to assessments, but I want to highlight the advantages of scales that are tied to learning progressions and that have relatively few discrete points linked to the progression.

There are, however, two caveats associated with this positive view of the impact of learning progressions on validity arguments. First, this approach rests on a number of assumptions: that the learning progression is clearly stated, with clearly differentiated levels of achievement; that the learning progression captures the full range (more or less) of what we want students to learn; and that the scale scores associated with particular score levels accurately reflect the kind of achievement defining that level. The last of these assumptions gets us into the methodology of standard setting, which is not all that satisfactory. However, assuming that the performance progression is clearly defined, and the test is explicitly developed to reflect the progression, the task of linking score levels to levels in the progression should be much easier and more robust than it is in many other contexts, where relatively general performance standards (e.g., basic, proficient, advanced) are imposed on existing tests.

Second, the interpretation of test scores is validated in terms of the learning progressions, and any additional implications that might be entertained would generally require additional analysis. For example, a strong case for the assumption that the score scale reflects the learning progression is not necessarily a strong case for a claim that the test scores (aggregated in some way) or the performance progression constitute a good basis for an accountability system. Both of the presentations pointed this out, and in particular, it is very difficult to justify these causal inferences inherent in accountability systems. The causal inferences are essential for accountability systems; if we are going to hold a school accountable for student learning, we have to assume that the school is the major causal element in student growth and that the impact of other construct-irrelevant (i.e., school effectiveness-irrelevant) factors is inconsequential.

A major weakness in validation practice is a tendency to validate some basic interpretation of the scores and then to casually extend the interpretation or to slide over to another interpretation. In logic, this kind of implicit shift is referred to as *begging the question*. One provides a strong (perhaps ironclad) argument in support of some proposition, X, but then moves on, assuming that one has proven some stronger proposition, which is, in some way, a more ambitious assertion than X. So we can build a strong

validity argument for an interpretation of test scores in terms of learning progressions, but we should not overstate what we have done.

In evaluating the reliability or precision of test scores linked to learning progressions, many traditional, norm-referenced models may not be appropriate, because these coefficients assume that the signals to be detected are differences between individuals in their levels of achievement. The precision of test scores interpreted in terms of learning progressions could be better described in terms of error/tolerance ratios, where the tolerance would be defined in terms of score differences between major levels in the progression.

In most areas in science and engineering, we evaluate precision in terms of whether the errors interfere with a reasonably accurate interpretation of observations or with appropriate decisions. That is, does the error cause a problem? Measurement error is acceptable if it is within the tolerance allowed by a particular application; if it exceeds the tolerance, it is too large. In using the test scores to assess growth in the learning progressions, the tolerance could be specified in terms of typical year-to-year growth, or in terms of the differences in the scores associated with different levels in the learning progression. Assuming that we want to be able to distinguish between different levels of the progressions, so that students are typically assigned to the correct level, we want the error to be small compared to the score difference between levels.

## **Value-Added Models**

For purposes of accountability, value-added models are conceptually better than plain achievement models like No Child Left Behind (NCLB). They at least try to take into account where students start. Nevertheless, I am skeptical about them for a number of reasons. First, they are very complicated, and we do not have a lot of experience with them. Being a New Yorker originally, I tend to be skeptical and am reluctant to accept something I do not fully understand. At some point we are going to have to explain these things to the teachers for whom high-stakes decisions are being made, to the school officials and school boards, to the parents, and perhaps to judges and/or congressional committees. So there is a benefit in keeping the models relatively simple if we can.

Second, some value-added models employ student change scores as the basic data, and this introduces several potential problems. First, change scores tend to be unreliable, and aggregating the change scores over a class or school can exacerbate the reliability problem, rather than reduce it. Value-added models that employ change scores run into all of the problems associated with vertical scaling. The use of score scales tied to learning progressions could conceivably help with this problem, but that is yet to be decided.

Third, if value-added models employ regression techniques, they run into another kind of problem; there are lots of subtle regression artifacts that could introduce bias into the results of these models, particularly regression to the mean. In comparing regressions across groups (e.g., classes or schools), regression effects can suggest differences that are not there and can mask differences that are there.

Fourth, there is potential bias associated with the choice of variables that are partialled out of consideration in the value-added models. Students are not randomly assigned to teachers or schools, and the systematic effects associated with such non-random assignment can be substantial and are very difficult to correct for. It may be that all of these problems have been taken care of in the value-added models, but I am quite skeptical, because the value-added models remind me a lot of the covariance models that were popular about 40 years ago. In the case of covariance analysis, my sense is that we decided that we did not know how to make the adjustments with confidence because if one does not have a complete set of variables to be used as covariates, one can get biases of various kinds. In that case, we seem to have given up on what initially seemed like a very promising approach to causal analysis based on correlational data.

## **Paradigm Shifts**

Betebenner and Linn talked about the desirability of a paradigm shift to an accountability system that is more descriptive and less punitive. This would be great, and we should all try to support this kind of shift as much as we can. It is, however, going to be an uphill battle, in part because paradigms are hard to shift. People want simple answers to complicated questions. When people can be engaged to look at the data, you can modify that somewhat, but getting people to look at the data is not easy to do on a sustained basis. It is a first step.

Changing the paradigm will be very difficult, but it might be possible to at least nudge the paradigm a bit in the direction of being more descriptive and less punitive.