Michigan Assessment Consortium
Improve learning.
Increase success.

Thinking About Improvement in Student Test Performance

James A. Gullen

October 2012

Our Context

Student achievement is at the heart of our educational systems. Quantifying and describing student achievement has proven to be a more complex task than many might have expected. For a number of years, accountability systems have described student achievement in terms of performance relative to a standard. This frame of reference has proven to be problematic for students who are very much below the set standard. For these students performance may improve but still not meet the threshold for the standard. Growth holds the promise of "giving credit" for improvements in student achievement even if the standard has not been met. Accountability systems also look to growth as a way to "correct" for demographic circumstances that are known to impact student achievement as measured on standardized tests (i.e., poverty). Teachers look to growth measures to provide information about student achievement within their classrooms ("How much more do my students know than when they started my class?")

"Quantifying and describing student achievement has proven to be a more complex task than many might have expected."

While the notion of growth in the surface seems like a relatively simple notion, it is actually quite complex. There are a number of different ways to think about student improvement that are current in the research. This paper will consider three primary ways of examining changes in student test scores. For some people, these methods are referred to interchangeably. This is problematic as it confounds the issue because valid interpretations of each are different. The requirements of the tests and/or scales are also different for the three methods. This short article will attempt to provide a brief explanation of each.

Pre-Post Testing

In pre-post testing, students are given an assessment before relevant instruction begins. Instruction takes place and then the exact same test is given a second time. Average post-test scores can be compared to average pre-test scores to get a feel for how a group improved. Individual student pre-test and post-test scores can also be reviewed to see improvement at the individual level. Three things are important to note: First, the pre-test and the post-test are the exact same test. Not just similar tests. Tests where the items have been reordered or the response choices scrambled do not count as the exact same test. Second, pre-post testing only speaks to improvement in student performance; it doesn't attribute that improvement to anything in particular. No causal relationship is validated. Finally, pre-post testing can be thought of as a criterion referenced viewpoint since one student's improvement doesn't depend on anyone else's test performance. Every student can improve. There don't have to be "winners" and "losers". Research will sometimes use pre-post testing along with research design to detect

causal relationships (treatment/control groups). In these cases, it is the research design that establishes the causal relationship. The pre-post test simply establishes the change.

Growth Models

Growth Models are similar to pre-post testing except that the tests can be different for the pre-test and the post-test. This does not mean that any two tests can be used. The tests used for administration before instruction and after instruction need to be "equivalent forms" or "equated". Essentially this means that while the tests are different they are measuring the same thing, on the same scale, with the same precision. This requires good test development with strong psychometric information to establish the "parallelness" of the measures. Reordering the items on a test or scrambling the response choices does not make an equivalent form for purposes of growth modeling. Like pre-post testing, one student's growth is not impacted by another student's growth. No attribution of what caused the growth is made using the growth model unless an appropriate research design is used in conjunction with the growth.

Value added models (VAMs) are the most complex of the three types of models. VAMs require tests with strong psychometric properties. The models used to estimate the value provided by each educator are very complex statistically. These models also make strong assumptions about the nature of the data that are fed into them to estimate these models. VAMs that are well estimated and fit the data well support causal inferences about changes in student achievement (i.e., being in teacher A's class caused 24% growth for his students and being in teacher B's class caused 56% improvement for her students). Value added models are not correct or incorrect. The fit of the model to the data is assessed and then determined if the model is "good enough." The higher the stakes associated with a VAM, the better the model needs to fit the data. If the tests used to provide data to the VAM don't adequately define the scale or if the data do not conform to the model's assumptions valid interpretation is compromised. Value added models are based on regression models, lots of regression models, actually. As such, VAMs are a normative frame of reference to look at the data. Some districts/schools/classes will be “above the mean” and some will be below (although lots of means are used in the models). The practical implication of this fact means that district’s/school’s/teacher’s “value added” is dependent on not only what went on in that particular group, but also what happened in the other district/schools/teachers that are included in the VAM.

“This descriptive material is not intended to convey that one approach to IBAs is better than another, or is to be preferred, just that they are different.”

Summary

In summary, quantifying improvement in student scores is an important topic in education. Research around this topic is vigorous and widespread. Thinking clearly about what we mean when we talk about student “growth” (as a generic term) will help us choose an appropriate model to use in attempting to quantify and interpret that data. Understanding the differences between these models, and being precise in our use of them, will help us move forward toward better models and make valid inferences from them.

"The mission of the Michigan Assessment Consortium is to improve student learning and achievement through a system of coherent curriculum balanced assessment and effective instruction."

The mission of the Michigan Assessment Consortium is to improve student learning and achievement through a system of coherent curriculum balanced assessment and effective instruction. We do this by collaboratively:

- Promoting assessment knowledge and practice.
- Providing professional development.
- Providing and sharing assessment tools and products.

www.michiganassessmentconsortium.org