Collegio Carlo Alberto

Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member

Kevin Lang

No. 143
April 2010

Carlo Alberto Notebooks

www.carloalberto.org/working_papers

Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member

Kevin Lang

Kevin Lang is Professor of Economics at Boston University, Boston, Massachusetts. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts, and Research Fellow, Institute for the Study of Labor (IZA), Bonn, Germany. His e-mail address is <lang@bu.edu>.

One of the potential strengths of the No Child Left Behind (NCLB) Act enacted in 2002 is that the law requires the production of an enormous amount of data, particularly from tests, which, if used properly, might help us improve education. As an economist, I have been appalled by the limited ability of districts to analyze these data. As someone who was first elected to the School Committee (called the School Board or Board of Education elsewhere in the United States) in Brookline, Massachusetts in May 1996 and served until May 2009, I have been equally appalled by the cavalier manner in which economists use test scores and related measures in their analyses.

When it comes to these newly available data on education, the idea that "measurement matters" can be interpreted in at least three ways, and each of these is correct. Measurement may be used as an adjective: this paper is principally concerned with matters of measurement. I have chosen to address this topic because the act of measuring changes behavior, which is another way in which measurement matters, and the conclusions we reach depend on how we measure outcomes, which is yet another way in which measurement matters.

In principle, tests mandated by the No Child Left Behind Act play several roles. They help guide instruction by, for example, helping districts discover areas that are not covered well in their curriculum. They are used in evaluation, primarily of schools and districts but increasingly of teachers, as well. Often they also serve as gateway tests for students, particularly as a requirement for high school graduation. My focus will be on the use of the data for evaluating performance by principals, by districts and other educational authorities, and by economists.

The kinds of data that we currently provide are very hard to interpret, and policy-makers, who lack statistical sophistication, cannot easily use them to assess progress. In some domains, most notably the use of average test scores to evaluate teachers or schools, the education community is aware of the biases and has sought better measures. The economics and statistics communities have both responded to and created this demand by developing value-added measures that carry a scientific aura. However, economists have largely failed to recognize many of the problems with such measures. These problems are sufficiently important that they should preclude any automatic link between these measures and rewards or sanctions. They do, however, contain information and can be used as a catalyst for more careful evaluation of teachers and schools. Perhaps more importantly, they can be used as a lever to induce principals and other administrators to act on their knowledge.

District Resources

One of the goals of No Child Left Behind is to increase the availability of data. Part of the implicit model underlying No Child Left Behind is that with improved information, parents will recognize good and bad schools. Principals will identify good and bad teachers. District administrators will identify weak and strong principals, and state administrators will recognize struggling school districts. Armed with this information, parents will choose with their feet, and the other actors will undertake the necessary reforms to improve education.

As an empirical economist I am, of course, sympathetic to the use of data, and as a school board member I pushed for more thorough evaluation of our programs. But the gap between the

rhetoric and the ability to use education data effectively is large.

Few school districts have the resources to analyze statistical data in even remotely sophisticated ways. In the early days of the Massachusetts Comprehensive Assessment System tests, I visited the Assistant Superintendent for Curriculum and Instruction who was anxious to use the testing data to help Brookline address its achievement gap. The state Department of Education had provided each district with a CD with the complete results of each student's MCAS test. In principle, it would be possible to pinpoint the exact questions on which the gap was greatest. The problem was that no one in the central administrative offices could figure out how to read the CD. I loaded the CD onto my laptop and quickly ascertained that the file could be read with Excel. Shortly thereafter, our Assistant Superintendent attended a meeting of her counterparts from the western (generally affluent) suburbs of Boston and discovered that Brookline was the *only* system that had succeeded in reading the CD. Districts have become somewhat more savvy about using data. A younger generation of administrators has more experience with computers, but relatively few would be able to link student report cards generated by the school district with SAT scores and the state tests.

Principals, district administrators and even state-level administrators generally begin their careers as teachers, and relatively few teachers have strong backgrounds in statistical reasoning. In my experience, the people who rise to senior administrative positions in public education are smart. They understand in a general sense that estimates come with standard errors attached, but faced with a report that last year 43 percent and this year 56 percent of black students in fourth grade were proficient in math, few could tell you whether with 75 students each year, the change was statistically significant.

When I stepped down from the school board, one of my colleagues joked that they could all go back to treating correlation as causality. In education policy settings, one repeatedly hears statements like: "Students who take Algebra II in eighth grade meet the proficiency standard in grade ten. We must require all students to take Algebra II in eighth grade." "Students taking math curriculum A and curriculum B get similar math SAT scores. The curricula are equally good." "Students who are retained in grade continue to fall further behind. Retention is a bad policy."[1]

School administrators may understand at some level that they are only looking at correlations, but almost none have the training to address the issue of causality, and faced with a correlation, they will often interpret it causally in the absence of evidence to the contrary. The capacity to address causality, weaknesses of various measures, and other strengths and weaknesses of statistics is very limited.

The Public Schools of Brookline recently recruited for a Director of Data Management and Evaluation. Although school board members generally are not (and should not be) involved in personnel decisions other than those involving the Superintendent, in this specific case the Superintendent asked me to participate in the candidate interviews. Many of the candidates held or had held similar positions in other districts. I asked each candidate how we could decide whether a math curriculum used by some, but not all, of our students was effective. Many of the candidates did not think of this question in statistical terms at all. Only one addressed the issue of selection—and we hired him.

1 See Nagaoka and Roderick (2004) and Jacob and Lefgren (2004) for sophisticated discussions of the effects of retention in Chicago and my discussion of these papers in Lang (2007, pp. 13-17).

Ordinality

School administrators and policy-makers are eager for help and have turned to economists and statisticians. Partially in response to these demands and partially because they were already developing relevant techniques, the research community has responded, but it, too, has been insufficiently cautious in interpreting the data.

Imagine a labor economist presenting a seminar in which she reported that the incomes of Xs have grown faster than those of Ys. Assuming that the audience did not instantly fall asleep, they would doubtless ask questions about the data. Were the earnings real or nominal? If nominal how, should they be adjusted? (Does the fact that the reported U.S. poverty rate is higher now than in the mid-1970s reflect an actual rise in poverty, or does it reflect inaccuracy in the Consumer Price Index used to adjust the poverty line each year?) Are the data top-coded, and does the top-coding affect the two groups differently? If the distribution of income differs for the two groups, is the mean the right metric for comparing incomes?

Now imagine that in the course of the seminar, this labor economist acknowledged that she did not actually measure income but only an unknown monotonic transformation of income. The codebook and questionnaires had been lost. All she actually knew was that the data were collected in categories and that higher codes represented higher incomes. The highest value observed in the data, 98, represented a higher income than did 97, but she did not know whether the distance between 96 and 97 was greater than, less than or equal to the gap between 2 and 3. Moreover, she was unsure whether the scales had changed over time. Nevertheless, she had blithely averaged the reported codes and then compared the changes in these averages. At the

bare minimum, this labor economist would have to show that her conclusions were robust to monotonic transformations of the scale. And even so, she would be lucky to survive the seminar without being either bloodied or ignored.

Yet when it comes to studies of education, economists often ignore these same concerns. Economic studies of education commonly proceed as if the intervals between scores always mean the same thing, as if top- and bottom-coding did not exist, and as if fourth- and fifth-grade test scores are really comparable. Psychometricians (analogies test: econometrician is to economist as psychometrician is to _____?) often devote considerable effort to equating test scales so that a scaled score of 150 on the fourth grade test represents the same level of achievement as a scaled score of 150 on the fifth grade test. However, there is no reason to believe that the resulting scale has equal intervals. We cannot say whether a student who receives a 120 in fourth grade and 180 in fifth grade has made more or less progress than one whose scores were 150 and 190.

Frequently economists will express discomfort with raw scores or other scores produced by psychometricians and transform the reported scores so that they have mean zero and variance one, but this practice is just a linear transformation of a variable with an unknown scale. Thus it, too, does not have equal intervals.

In practice, as long as we are concerned with differences in true averages, I expect that relying on the average of an arbitrary scale will not lead us too far astray. The ranking of schools with very different outcomes is likely to be invariant to the choice of scale: that is, the cumulative distribution functions probably do not cross. Schools that are similar using one choice of scale may differ according to another, but as long as we are careful not to put too much

weight on small differences, we will probably not end up describing test scores as much higher at one school than at another when the opposite would be true if we had chosen a different but plausible scale. But as soon as we are interested in measuring differences in progress over time, such issues are likely to become important.

As an example, consider evaluating whether a state is meeting one of the principal goals of No Child Left Behind legislation—reducing the measured achievement gap between blacks and whites. Given data on test scores, surely this question is straightforward to answer?

Table 1 shows the distribution of scores on the 10th grade Massachusetts Comprehensive Assessment System statewide mathematics exam for blacks and whites in 2001, the first year for which passing the 10th grade MCAS was a graduation requirement, and 2008. For the moment, let us ignore all of the other concerns we may have about the tests and the results (for example, test score inflation, whether the tests capture what we want students to learn, adjusting for differential dropout rates) and treat the test scores as good measures of student learning. Has the test score gap narrowed?

If we believe the scale implicit in the No Child Left Behind legislation—essentially that proficient or higher is worth 1; anything below proficient is worth 0—the answer is "somewhat." The proportion proficient or advanced increased by 32 percentage points (from 16 to 48) among blacks compared with 27 percentage points (from 51 to 78) among whites. Given the number of students taking the exams, this difference is undoubtedly statistically significant. Still, the progress is modest compared with the 30 percentage point gap that remains.

But we might interpret the results very differently if we were primarily concerned with raising the bottom of the distribution. Massachusetts requires students to get at least "needs

improvement" on the 10th grade tests in order to graduate. Here the gains of black students relative to white students are quite large. The proportion scoring at least needs improvement has increased by 32 percentage points (the proportion failing decreased by 32 points) among blacks compared with only 13 percentage points among whites. On the other hand, if we are worried that there are not enough blacks performing at very high levels, we might focus on gains at the advanced level and conclude that blacks had made less progress than whites.

Where we come down on whether the achievement gap has narrowed will depend on the scale we use. Faced with the data in Table 1, many education policy-makers would note the complexity of the question. But if education policy-makers are told the result based on an arbitrarily chosen scale—like "share of students reaching proficiency" or "average scores for black and white students"—few would recognize that they are hearing only one possible response to a complex question.

Unfortunately, this problem does not have a simple solution. Suppose you are told that a policy intervention raised the wages of poor whites by $10,000 more than those of poor blacks and of rich whites by $10,000 more than of rich blacks, and you know the proportions of poor and rich blacks and whites. In most cases, you still do not have enough information to determine whether the wages of whites or blacks grew faster. If a much higher fraction of blacks than of whites are poor, and the wages of the poor grew faster than the wages of the rich, black wages could still have risen faster than white wages even though they rose less rapidly within each sub-group. Since we do not know the value of going from a score of 120 to 160 relative to the value of going from 220 to 260, our position is like that of the program evaluator who only knows the relative salary gains within group.

As long as blacks are over-represented at the lower end of the achievement distribution, whether the gap has increased or decreased will depend on how much weight we put on improvements at that end of the distribution. To avoid using an arbitrary scale requires us to create an interval scale from the test scores either by being explicit about our values or by linking the scale to some future outcome such as high school graduation or college attendance, neither of which is easy. In practice this means that we are unlikely to be able to make strong statements about whether we are closing the gap if both test score distributions are improving. One solution is to ask whether the conclusion is robust to reasonable transformations of the scale. A more powerful solution is to focus on different questions: at what points in the test score distribution are we making progress, and are blacks and whites making similar progress at these points in the distribution?

Categories and Performance Standards

The choice of performance categories like what should qualify as "proficient" or "basic" or "failing" is arbitrary although, as Koretz (2008, p. 183ff) also stresses, this does not mean it is capricious. In the best of cases, proficiency and other performance categories are determined by experts using their best judgment. Leaving aside political pressure to lower the proficiency standard to make it easier to meet No Child Left Behind standards, I have no reason to think that this process is conducted with anything other than good faith.

But the outcome of choosing performance categories for tests, if taken a face value, can sometimes lead to improbable conclusions. One day, I received a phone call from a candidate for

local office who was concerned that the mathematics performance of black students in Brookline deteriorated sharply between grades six and eight. I was pretty sure I knew what the problem was, and a quick check confirmed my suspicion. In 2008, 56 percent of Massachusetts sixth graders were found to be proficient or advanced in mathematics, compared with 49 percent of eight graders and 72 percent of tenth graders. Now we cannot fully rule out the possibility that the quality of mathematics education in Massachusetts is catastrophic in grades 7 and 8, and then recovers in high school. Moreover, certain additional factors may inflate tenth grade proficiency rates. But it seems much more likely that the eighth grade proficiency standard is high relative to those in the other two years. Since in my own Brookline school district, a higher proportion of blacks than of whites is near the proficiency cutoff, the higher standard in the eighth grade manifested itself as a large increase in the achievement gap.

This is a specific example of a general problem with comparing performance on the basis of categories. On most tests, more whites have high test scores and more blacks have low test scores. If one chooses a high cutoff, there will be more whites than blacks just below the cutoff, and similar improvement in the underlying test scores for both groups will push more whites above the cutoff. Similarly, if one chooses a low cutoff, more blacks will be just below that low cutoff and more will cross the cutoff as test scores improve for both groups.

This is illustrated in figure 1. It shows the distribution of test scores for two groups. The first (dotted) vertical line represents the cutoff between “unsatisfactory” and “proficient.” The figure has been drawn so that only members of the low group fall into the unsatisfactory performance category. Therefore any rightward shift in the distribution of its test scores will reduce the “proficient or better” gap between the two groups even if the distribution of high

group test scores improves a great deal more. The second (large dashed) vertical line in figure 1 shows the cutoff between "proficient" and "advanced." The figure has been drawn so that no member of the low group scores close to the advanced level. Similar modest improvements in the two test score distributions will increase the "advanced" gap.

For some purposes, the arbitrariness of the performance standards is unimportant. The black-white achievement gap exists regardless of scale. But if a school board wants to know whether the achievement gap narrows or widens between sixth and eighth grades, the use of performance standards is problematic and the use of performance standards that are inconsistent across grades is extremely problematic.

Value-Added Measurement

Currently the No Child Left Behind Act assesses schools and districts primarily on the basis of students' absolute performance on the state tests. But critics argue (correctly) that student performance measures not only school quality, but also other inputs such as parental support and the skills that students bring with them to school. So-called "status measures" of absolute performance are much easier for districts with advantaged students to satisfy than they are for those with disadvantaged students.

One solution is to adjust the status model for characteristics of the students. However, such "adjusted status models" are likely to overcompensate for differences in the conditions facing schools. Regression methods effectively ask whether a school is better or worse than the average of similar schools. But the premise underlying the No Child Left Behind Act is that poor

children are often “left behind” in underperforming schools while more advantaged parents move their children to higher-performing school districts. If this premise is correct, then controlling for “middle-class parents” also controls for some of the difference in school quality since better schools attract more middle-class parents. We would not compare the incomes of two people by asking how much they earned relative to the housing prices in their neighborhood. We would guess that someone with income somewhat below the average in a neighborhood with high housing prices earns more than someone with an income somewhat above the average in a low-price neighborhood. Similarly, we would expect that a school in an upper-middle class neighborhood that performs only somewhat below the average for such schools is a better school than a school in a poor neighborhood that performs only somewhat better than the average of such schools. If not, there would be little justification for the NCLB legislation.

Many critics of status measures argue for the use of growth or value-added measures of school performance. The simplest approach is a model in which we measure school quality by the change in student performance between year t and t+1. The models used in practice are typically much more complicated and sophisticated, but the basic issues about scaling remain.

Consider the following simple example. Massachusetts computes a “composite performance index” that awards 100 points for a student achieving proficiency or better, 75 points for those in the upper half of the “needs improvement” category, 50 points for those in the lower half of this category, 25 points for those in the upper half of the warning/fail category and 0 for those in the lower half of this category.

We can generate a pseudo-growth or gains-score model by comparing the composite scores index for a district in grade *g* in year *t* with the corresponding index score for that district

in grade *g-1* in year *t-1*. Of course, if we wanted to do this "right," we would have to take account of inter-district mobility, but this factor is unlikely to be important for the point I am making. I take 2007 and 2008, because these are the most recent years currently available, and 5^{th} and 6^{th} grade mathematics because the distribution of performance on these two tests is fairly similar, suggesting no major problems with differing standards.

In 2007, 5th graders in Holyoke had a composite score index of 46.0 in math while the 6th graders in 2008 had a corresponding score of 50.3. In 2007, 5th graders in Winchester had a composite score index of 98.1, but 6th graders in 2008 had a corresponding score of only 94.0. Given these status measures, it is not surprising that Holyoke is widely regarded as one of the worst school districts in the state and Winchester as one of the best. But the gains scores tell a very different story. Using changes in the composite scores index as our interval scale, Holyoke sixth graders gained 4.3 points while their counterparts in Winchester lost 4.1 points.

The Holyoke-Winchester comparison is not an aberration that I was able to discern by poring through the data. In fact, among Massachusetts school districts, the correlation between the gains score and the average score over the two years is approximately zero.[2] The problem is that if we rescale each year, we do not really have a measure of learning, we have a measure of relative performance. If relative performance is persistent, then most of the year-to-year change relative to the state mean is random fluctuation, a point made in much greater detail in this journal by Kane and Staiger (2002). Yet economists routinely rescale tests each year.

The use of relative measures also has important implications for economic research. A common norm in the economics profession is to take whatever test score is available and rescale

[2] Not surprisingly, given regression towards the mean, the gain is negatively correlated with the 2007 score and positively correlated with the 2008 score.

it to have mean zero and variance one, and this approach has important effects on the conclusions we draw from the data. There is a pretty strong consensus among economists and statisticians that "teachers matter"—that is that some teachers contribute more to student learning than do others. But there is also a pretty strong consensus that teacher effects fade out quickly (McCaffrey et al., 2004; Lockwood et al., 2007; Jacob, Lefgren, and Sims, 2008; Kane and Staiger, 2008). Given the way we rescale scores, how could it be otherwise?

Consider an extreme case. Suppose that each student's learning was determined solely by his teacher. Given appropriate measurement, then learning will be the sum of everything that student has learned from a series of teachers over time. To keep the example simple, assume that teachers are randomly assigned to students, that the variance of teacher contributions to learning is constant across grades and that the econometrician is provided with a perfect measure of learning. Given these assumptions, the variance of the measure the econometrician receives grows linearly with years in school. But the econometrician wants the variance to be constant over time. He therefore divides the second year scores by the square root of two, the third year scores by the square root of three, and so on. Then the econometrician regresses the rescaled scores using a dummy variable for each teacher.

Recall that using the perfect measure of learning, student learning was simply the sum of teacher effects. But with the rescaling, for second graders, the measure is sum of these teacher effects divided by the square root of two. Therefore, any contribution of a first-year teacher to learning declines to about 71 percent of its original effect in year two, 58 percent in year three and so on. In this example, fadeout will seem slower for teachers in later years. The extent of the fadeout will be determined by the scale-specific choice of how much we allow variance to

increase over time and can be reversed if we allow it to increase sufficiently.

Despite such concerns, in the policy world the use of value-added measurement for assessing educational performance gains support from its scientific aura. It appears possible to determine scientifically which schools and/or teachers are contributing the most to student learning. In contrast with models based on students' uncorrected test scores, measuring value-added appears capable of removing the effects of other factors. Yet, in some ways the opposite is true. Most principals, school district leaders and policy-makers can make ad hoc and admittedly imperfect adjustments for the socio-demographic composition of classrooms and schools. However, the complexity of even relatively simple models of value-added makes it difficult to determine the nature of their bias. Few skilled statisticians, let alone educators, could accurately predict the bias of a particular model.

For example, suppose that there are two types of students, weak and strong, who are perfectly identifiable by a pretest on which they score -1 and 1. As reported in the first panel of Table 2, test scores rise by two when weak students are matched with good teachers and by one when they are matched with poor teachers. The corresponding values for strong students are six and three. There are two teachers of each type, one of whom is matched with two strong and one weak student and the other of whom is matched with one strong and two weak students. In this setting, the second panel of Table 2 shows that simply examining mean test scores will favor the teachers with a larger share of strong students.

One way to calculate value-added is to calculate each student's predicted score based on the pretest score and then calculate the average difference between the actual and predicted scores. Alternatively, we can regress student scores on their pretest scores and teacher fixed

effects. Economists tend to favor the latter approach while statisticians tend to favor the former. The example has been set up so that they give the same answer. I show the first approach here since the calculations are easier to present.

Half of the weak students have poor teachers and half good teachers. So the average score of a weak student is .5. Similarly, the average score of a strong student is 5.5. Therefore an average teacher with two weak and one strong student is expected to have total test scores of 6.5 while one with two strong and one weak student is expected to have total test scores of 11.5. Value-added is calculated as the difference between the actual and predicted scores.

The resulting value-added scores, summed over all three of the teacher's students, are shown in the third panel of Table 2. It turns out that the teacher with the lowest estimated value-added is the poor teacher matched with two strong students and the one with the highest estimated value-added is the good teacher matched with two strong students. There is no consistent bias favoring teachers with strong or weak students in this example, but the approach does not rank teachers correctly. And it is easy to generate transformations of the later test scores that reverse the ordering within each pair of teachers (for example, leave the pretest scores the same, but rescale all later test scores by multiplying by 3.9).

The point is not that value-added measurement inevitably gives the wrong answer. In this example, there is a more complex model that will give the "right" answer. That model allows the value-added of the teacher to depend on the nature of the student he teaches. There is also a transformation of the scale that makes the value-added of each teacher the same for all students. With this scale, the two standard approaches also rank teachers properly.

Instead the point is that even in very simple settings (four teachers, twelve students, two

types of teachers, two types of student, perfect pretests), the nature of the bias is extraordinarily difficult to anticipate. In more complicated settings, we simply cannot know whether we are over- or underestimating a particular teacher's effectiveness. In contrast, with status models, we often have a good sense of the nature of the bias.

Jacob and Lefgren (2008) have argued that principals do not correct sufficiently for student characteristics and instead put too much weight on test scores. They reach this conclusion by examining the relation between their estimate of value-added and principals' assessments of teacher quality. They find that while their value-added measure predicts principals' assessment, conditional on value-added, principals give higher ratings to teachers whose students have higher absolute test scores. The conclusion that principals rely excessively on test scores is only merited if the divergence between true quality of teaching and estimate value-added is independent of test scores. This condition is satisfied in the example given here, but it need not be in general.

To see this, consider the following example. Suppose there is a third category of stellar students and an additional good teacher who teaches all the stellar students and one weak student. Because she teaches all the stellar students, they all have the same outcome, and their high pretest score perfectly predicts this outcome. The teacher's value added for these students is 0. She does get value-added credit for the one weak student, but not as much as the other good teachers get for their strong students. Of course, principals may be poor statisticians, but evidence that principals rely "excessively on status scores" may not demonstrate that they are poor statisticians, but rather that economists lack adequate models.

Using Test Scores to Evaluate Schools and Teachers

There are a multitude of reasons to be cautious about using test scores to evaluate teachers and schools, from the risk of narrowing the curriculum to the absence of tests in many subject areas. The discussion in the previous section adds to the case against linking rewards and sanctions automatically to some measure of performance. However, I do not wish to imply that there is no information in estimates of value-added or status measures. Teachers with both very low value-added as measured by a sensible model and low mean student test scores are likely to be poor teachers. Those with high value-added and mean scores are likely to be good.

I do not mean this as a theorem, but rather as a pragmatic approach. It is expensive and time-consuming to evaluate teachers and schools, and we should target our resources to those cases where more careful review is likely to reveal weaknesses. On average, teachers whose students perform poorly are weaker teachers than those whose students perform well, but certainly many good teachers have students who perform poorly. Similarly, teachers whose students perform less well than would be expected on the basis of those students' past performance are, on average, weak teachers, but given the inevitable biases in value-added measures, many teachers with low measured value-added will be good teachers. Since status and value-added are imperfectly correlated, poor performance on both measures is a clearer indication of low quality than is poor performance on a single measure. In the past I argued that when forced to choose among imperfect models, we should deliberately choose those that are biased in favor of teachers of disadvantaged students to offset the bias in status models, but I confess that I have no idea how to do that.

Since both measures are subject to error, even used jointly they should not lead to automatic sanctions but rather to more careful investigation. Fortunately, as discussed briefly in the previous section, there is good reason to have confidence in principals' and administrators' subjective evaluations.

If principals and other administrators are doing their jobs, they already know who the best and worse teachers are. One year, Brookline administered one form of the Iowa Test of Basic Skills in the fall to third graders and the state administered a different form of the same test in the spring. I calculated the mean gain score for each teacher and asked the Assistant Superintendent for Curriculum and Instruction to tell me which teachers would have the highest and lowest gain scores. She got the group of highest teachers exactly right and missed only one of the lowest. I would expect building-based administrators to do even better. More formal analysis confirms my experience (Murnane, 1975; Jacob and Lefgren, 2007).

Therefore, while status and value-added measures may sometimes provide them with information, the real issue is not helping administrators determine which teachers are poor and which good, but to give administrators the power and incentives to act on this knowledge. In Massachusetts, teachers who have worked in the same district for three years are entitled to "professional teacher status" in that district. Teachers without professional teacher status may be fired without cause. Firing a teacher with professional teacher status requires "good cause." With these rules, retaining a teacher beyond the third year makes that teacher much more permanent. If test scores are to play a role in this decision, then the granting of professional teacher status will be based on only two years of test scores. For teachers who are in their first job, the learning curve over the first three years is quite steep. Ignoring evidence from third year performance

would be problematic. One option would be to make the cutoff for professional teacher status either the greater of three years in the system or five years as a teacher. This policy would give districts that primarily hire inexperienced teachers more time to evaluate them. While professional teacher status is often referred to as “tenure” and some states use this term, at least in Massachusetts, it is possible to get rid of experienced teachers. The Massachusetts courts have ruled that “good cause” requires that a dismissal be in good faith and relevant for creating and maintaining an efficient school system. By this standard, being ineffective is a “good cause” for dismissal. But establishing that a teacher is ineffective requires careful observation and evaluation. And the process must be defensibly fair – a principal who determines that a teacher appears to be ineffective and then focuses evaluation efforts only on that teacher is likely to find that a court will not find this evidence adequate to dismiss the teacher unless the principal has also conducted similar evaluations of other teachers.

Evaluating all teachers consistently is time-consuming. Principals who succeed in dismissing tenured teachers (or even teachers without tenure) risk angering the remaining teachers. In many school districts, principals move frequently and therefore are unlikely to reap significant benefits from getting rid of weak teachers. It is not surprising that many, perhaps most, take the easy path and give only good evaluations.

It is here that using test scores can be useful. Having both low value-added and low mean scores is an indicator that a teacher is ineffective. It should be possible to make this outcome an automatic trigger for further evaluation. Some teachers targeted by this trigger will turn out to be effective in the principal’s judgment, but most should not. If principals were required to justify their decision to retain teachers identified in this way, the incentives would be shifted. District

administrators could question principals who consistently chose to retain teachers identified in this manner, and knowing this, principals would have more incentive to evaluate teachers carefully.

Conclusion

Standardized tests generate a litany of concerns about the tests themselves, the subject material coverage, and the measures derived from them. While I have sidestepped many of the broader issues in this paper, I have emphasized that our ability to draw strong conclusions from test results is limited by measurement issues that we are unlikely to overcome soon, if ever. However, I do believe that we can use tests as a trigger for further investigation. Teachers and schools with both low average test scores and low estimated value-added are likely to be weak and, given limited resources, this group is the obvious place to target. Since tests are only very imperfect measures of the desired outcomes of education, we should avoid any automatic consequences for teachers or schools based solely on measures derived from tests. However, in the case of teachers, we know that in most cases, qualitative investigation by principals will confirm the statistical analysis. I expect that the same will prove true if qualified reviewers inspect schools.

The challenge for the next stage of education reform is therefore not to find better statistical methods for identifying low-performers but to provide administrators with the necessary resources to understand the statistical information and the ability and incentives to act on both statistical and qualitative information.

Acknowledgements

I owe an unusual debt to Superintendents Bill Lupini, Dick Silverman and the late Jim Walsh and to many senior staff of the Public Schools of Brookline with whom I worked for many years. My colleagues on the Brookline School Committee not only tolerated my outbursts on matters statistical but helped me refine my thinking about the role of data in informing education policy. Members of the National Research Council Board on Testing and Assessment and its panels on Incentives and Test-Based Accountability in Public Education and on Value-Added Methodology in education and the NRC staff will, I hope, recognize their enormous influence on my thinking. I am grateful to Rick Hanushek, Dick Murnane, Derek Neal and the editors of this journal for helpful comments. This paper was written while I was a visiting fellow at the Collegio Carlo Alberto and at the University of New South Wales. I am grateful to both institutions for their hospitality. Any errors of fact or judgment are, of course, solely my responsibility.

References

Jacob, Brian A. and Lars Lefgren, 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *Review of Economics and Statistics*, 86:1 pp. 226-44.

Jacob, Brian A. and Lars Lefgren, 2007. "Principals as Agents: Subjective Performance Assesssment in Education." *Journal of Labor Economics*. 26:1 pp. 101-136.

Jacob, Brian A., Lars Lefgren and David Sims, 2008. "The Persistence of Teacher-Induced Learning Gains," National Bureau of Economic Research Working Paper 14065, June.

Kane, Thomas J. and Douglas O. Staiger. 2002 "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, 16:4, pp. 91-114.

Kane, Thomas J. and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, National Bureau of Economic Research Working Paper 14607, December.

Lang, Kevin, 2007. *Poverty and Discrimination*, Princeton, NJ: Princeton University Press.

Lockwood, J. R., Daniel F. McCaffrey, Louis T. Mariano, and Claude Setodji, 2007. "Bayesian Methods for Scalable Multivariate Value-Added Assessment," *Journal of Educational and Behavioral Statistics.* 32, pp: 125 - 150.

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton, 2004. "Models for Value-Added Modeling of Teacher Effects" *Journal of Educational and Behavioral Statistics*, 29:1, pp. 67-101.

Murnane, Richard J. 1975 *The Impact of School Resources on the Learning of Inner City Children*, Ballinger: Cambridge, MA.

Nagaoka, Jenny and Melissa Roderick, 2004. "Ending Social Promotion: The Effects of Retention," Consortium of Chicago School Research.

Figure 1
Illustrative Example: Performance Categories

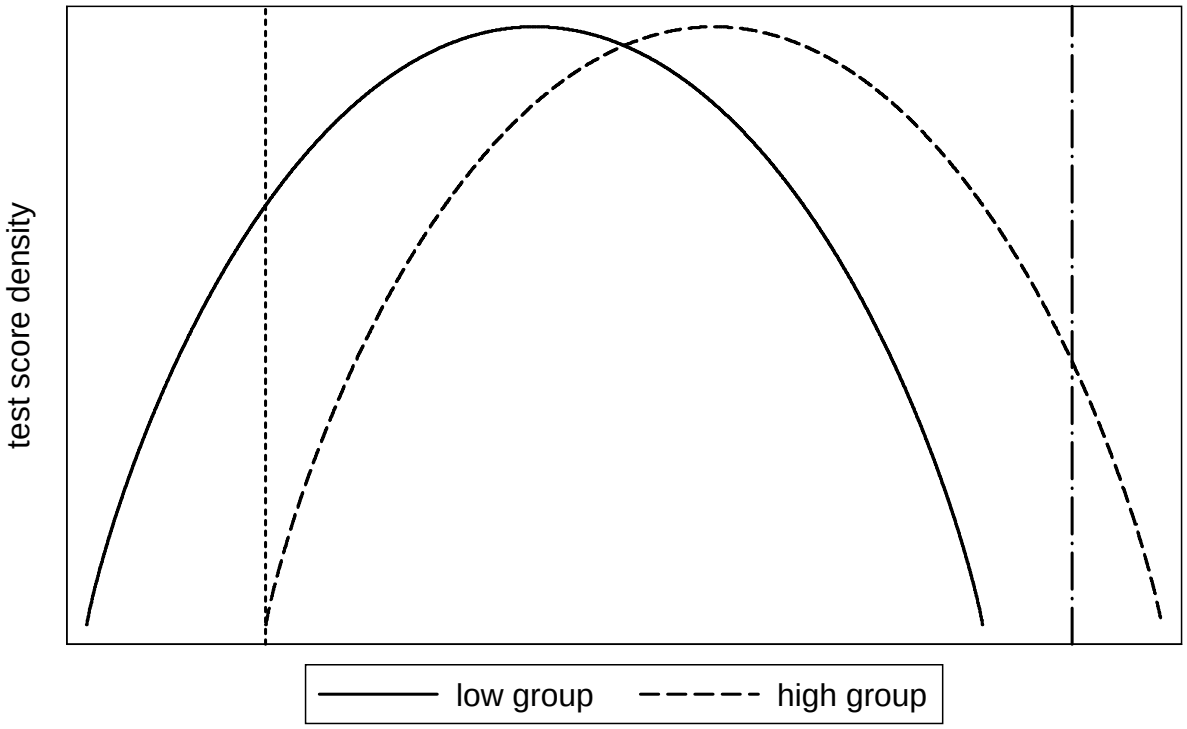

Table 1

Performance on 10th Grade Massachusetts Comprehensive Assessment System (MCAS) Exam in Mathematics, by Race

	Advanced	Proficient	Needs Improvement	Failed
White				
2008	48	30	16	6
2001	21	30	31	19
Difference	27	0	-15	-13
African American/Black				
2008	20	28	33	19
2001	3	13	32	51
Difference	17	15	1	-32

Rows sum to 100. Each entry is the percentage of the group falling into that performance category in the given year.

Table 2
Identifying Good and Poor Teachers from Test Scores: An Example

Test scores: Pretest plus gain equals final score

	Test score with good teacher	Test score with poor teacher	Mean
Weak student	-1 + 2 = 1	-1+1 = 0	.5
Strong student	1 + 6 = 7	1 + 3 = 4	5.5

Classrooms scores: Summing scores across students

	One weak student, Two strong students	Two weak students, One strong student
Good teachers	1 + 7 + 7 = 15	1 + 1 + 7 = 9
Poor Teachers	0 + 4 + 4 = 8	0 + 0 + 4 = 4

Value-added scores: Summing across students

	One weak student, Two strong students	Two weak students, One strong student
Good teachers	15 – (.5 + 2*5.5) = 3.5	9 – (2*.5 + 5.5) = 2.5
Poor teachers	8 - (.5 + 2*5.5) = -3.5	4 - (2*.5 + 5.5) = -2.5