

If Wishes Were Horses: The Reality behind Teacher Quality Findings

By Kate Walsh
President, National Council on Teacher Quality

An exciting proposition is making the rounds of the education policy circuit. Generated by recent findings on the importance of teacher quality, it suggests assigning great teachers five years in a row to a class of disadvantaged children could overcome the learning deficits of children who start out and then fall further behind in school over successive years. This feat—until now an educator’s fantasy—would be a hopeful breakthrough for a profession that for decades has assumed that schooling has little chance of overcoming a child’s family background.

The research behind this “five-year proposition” is solid, forged by some of the field’s most respected education economists, including Eric Hanushek. Their work over the past decade has provided compelling evidence that schools and, in particular, teachers have a major impact on children.¹ An economist from Tennessee, William Sanders, developed the statistical methodology that led to these findings, making it possible to measure fairly the effectiveness of an individual teacher over time. Sanders’s most noteworthy finding is that students assigned to a good teacher

for three years in a row will score 50 percentile points higher on tests than students assigned to weak teachers over the same period.²

These findings have been met with a feeding frenzy in education circles. They have re-energized a demoralized profession in sore need of affirmation and provided great sound bites for those pushing for across-the-board pay increases. But amid all this frenzy, very little discussion has taken place about what the teaching profession has to do to convert these findings from theory into practice. In the meantime, family environment continues to swamp the current impact of schooling.

In fact, not much evidence suggests that the education community is aware of some of the dramatic changes needed or the challenges that must be surmounted to implement the five-year proposal. Concerned that euphoria may be overtaking reality, a small handful of academics, especially Vanderbilt economist Dale Ballou, have tried to temper expectations. But this message of moderation has been delivered softly because these same economists worry

* Author Note: This working paper was prepared for the forum “Measuring Child Well-being,” on March 28, 2006, co-sponsored by Foundation for Child Development and the Brookings Institution. This paper has not been through a formal review process and should be considered a draft. This paper is distributed in the expectation that it may elicit useful comments and is subject to subsequent revision. The views expressed in this piece are those of the authors and should not be attributed to the staff, officers or trustees of the Brookings Institution.



1. Eric A. Hanushek, John F. Kain, Daniel M. O’Brien, and Steven G. Rivkin. *The Market for Teacher Quality*, National Bureau of Economic Research, February 2005 found “substantial differences in teacher quality when put in the context of student achievement growth. This implies that a one standard deviation increase in teacher quality raises standardized gain by 0.22 standard deviations. Since these quality variations relate to single years of achievement gains for students, they underscore the fact that the particular draw of teachers for an individual student can accumulate to huge impacts on ultimate achievement.”
2. William L. Sanders and June C. Rivers, *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement*, University of Tennessee Value-Added Research and Assessment Center, November 1996.

Robbing Peter to Pay Paul

2

that focusing too much on the challenges could result in a return to the status quo, making it unlikely that the overall quality of the nation's teachers will be improved on the scale needed. No obstacle, they assert, should stand in the way of the push to make better policy decisions about the teaching profession. In particular, schools must put an end to the obsolete notion that teacher compensation, determined by a uniform salary schedule, should be immune from market forces.

Robin Hood Comes Up Dry

This much is certain: Given the current pool of teachers and the way schools now operate, the chance of a poor child getting assigned a great teacher five years in a row through chance is negligible. This fact explains why Hanushek's five-year theory on closing the achievement gap has to be a conscious policy decision. A whole set of problems make achieving the five-year solution unlikely under natural conditions: the current pool of teachers, the distribution of teachers, how teachers get assigned to classes, and the distribution of poor children.

First, based on the data now available, only one in seven teachers meets the standard of effectiveness necessary to produce big learning gains each year. In other words, a child has a roughly 15 percent chance of being assigned to a great teacher in any given school year. Those are not impossible odds—but that's only one year. What are the odds of a child getting randomly assigned to a great teacher two years in a row? And if Hanushek is right, how about five years in a row? The odds are in 1 x 75, or 1 in 17,000.

But even these odds are better than the odds of a disadvantaged child being assigned a great teacher. The 1 in 17,000 odds assume that great teachers are evenly distributed across all schools and that high-poverty schools have as many great teachers as low-poverty schools. To the contrary, the view of most policy groups, teacher

unions, and policymakers is that schools serving poor children have the worst teachers and therefore are unlikely to have the same proportion of great teachers as wealthier schools. Certainly, the credentials of teachers in high-poverty schools are not as good as those of teachers in other schools. However, the research on the distribution of effective teachers is less clear. Evidence from a couple of studies (one by Hanushek and his colleagues in 2005 and another by Brian Jacobs and Lars Lefgren in 2006) indicates that schools serving poor children may house as many teachers capable of producing these large gains as schools serving more affluent children. Interestingly, their research shows far more variation of teacher quality within a school than among schools.

Even if poor schools do have as many great teachers as other schools, classroom assignment remains a problem. Teacher assignment within a school is frequently not random because it's rarely a secret who the great teachers are. Parents know who they are, and parents are not a silent bunch when it comes to doing what's right for little Suzie or Johnnie. More active, better educated parents will insist that their children be assigned to the best teachers. Some principals go along in order to keep the peace. Nor are effective teachers evenly distributed among grades. Principals routinely assign their better teachers to grades in which children will be tested, making it less likely that other grades have the depth of talent to do the job.

Finally, poor children are not distributed evenly across cities or towns. For all these reasons and more, the chance that nearly all of the teachers (rather than one in seven) in a high-poverty school could be superheroes is nil. For the purposes of this argument, let's assume that the current system produces genuinely great teachers roughly 15 percent of the time, though the actual percentage varies more than that suggests. Staffing a high-poverty school

with, say, 28 great teachers would require shifting the three best teachers from eight other schools. A Robin Hood policy of pooling all of a school district's scarce teaching talent into a few sites is politically untenable and probably not all that practical given that they'd likely be taking good teachers from schools that are not as poor but still plenty poor.

A 50-50 Proposition

The current system falls well short in providing all children with a fair and equitable education. Compounding the problem are significant methodological problems that prevent us from accurately identifying great teachers. The ability to identify great teachers before the school year begins and after it ends is crucial for both assigning teachers to the classrooms where they are needed and aligning compensation with performance.

To begin, the research findings of both Hanushek and Sanders are "ex-post" findings. Neither researcher conducted an experiment to identify a group of highly effective teachers ahead of time and then study their performance over the next several years to see what kinds of gains they produced. Instead, they studied many years of data from school districts after the fact, combing the records for evidence of teachers who had produced high gains in any given year.

Why is this distinction so important? Because it means that the researchers did not identify a cohort of effective teachers ahead of time and then find them to be effective year after year. It turns out that there is little stability from year to year—according to these value-added datasets—in who is or is not effective. Ms. Jones might be a highly effective teacher one year and one of the worst teachers in the school the next. The correlation between a teacher's performance one year and the next can be as low as 0.5. Given this instability, school administrators can't be confident that a particular teacher does

or does not deserve a performance bonus. Nor can they confidently decide whether a teacher who was enormously effective one year should be assigned to the school's most disadvantaged children the next.

These ups and downs in the data are troubling because they do not mesh with our own experiences of good teachers. We know certain teachers to be generally and continuously effective, not effective one year and horrible the next. The reason for all this bouncing around may not have much to do with a teacher's actual performance. Instead, the issue may be the limitations researchers face in measuring teacher effectiveness. When researchers measure huge impacts of teachers on student achievement, they are actually capturing all of the influences in the classroom—with the fair presumption that teachers are the dominant influence. Researchers have yet to figure out how to isolate the effect of the teacher from these "non-teacher classroom effects," such as the influence of other students on a child's ability to learn. These other classroom effects turn out to be quite "noisy"—not just literally but statistically, making it hard to measure accurately a teacher's true effectiveness. This noise might explain the instability of teacher performance—or not. The possibility remains that teacher performance fluctuates from year to year a lot more than anyone thinks. No one really knows.

There are ways to get around the instability of the data, but all of the available solutions require observing teachers over multiple years. Sanders accommodates this problem in his value-added instrument by averaging a teacher's performance over three years. Using a three-year average of student test scores, the correlation over time is much higher, making it a fair way to measure teacher performance. However, for the three-year average to work—under Sanders' methodology, at least—teachers have to stay in roughly the same assignment. Since many teachers do not stay

Robbing
Peter to
Pay Paul

4

in the same assignment that long, this requirement for identifying great teachers proves to be largely impractical.

While newer value-added methodologies may help address these challenges, the need to have multiple years of data throws up a serious obstacle to states or districts that hope to reward teachers for producing strong student gains in a given school year. Districts looking to reward teachers for student learning would have to include more than one valid measure. One option would be to test students more than once in a school year. However, the mere mention of more tests could incite riots from a test-weary profession.

Further, teachers don't like the idea of judging performance and compensation based on a test because they suspect it can be unfair. They oppose being judged by their students' performance on a single schoolday (or week). It would be an interesting proposition for a district to consider a distributed test strategy, with more tests of shorter length, scattered throughout the year, reducing the annual stress the current tests produce. Not only would multiple measures yield relatively stable assessments of teacher performance, but teachers would have much more opportunity to make midcourse corrections with more data. No matter how many tests are given, however, annual observations by the school principal or senior faculty, trained in how to administer the evaluation reliably, should always play a prominent role.

The Wobegon Effect Does Not Apply Here

When teacher performance is plotted on a graph, roughly 15 percent are decidedly more effective than the rest. Undoubtedly, an equal number are decidedly horrible, producing markedly low student gains compared to their colleagues. Most teachers' performance huddles around the middle. Statistically, teachers from about the 15th to the 85th percentile are nearly indistinguishable from one another. Econo-

mists are reluctant to make further distinctions among average teachers because doing so would require accepting a higher margin of error than is recommended for statistical modeling.

Actually, being able to distinguish between the top 15 and the bottom 15 percent is sufficient for many purposes, particularly for deciding who should be on probation and who should be awarded performance bonuses. But there is yet another catch: The top 15/bottom 15 distinction is apparent when analyzing student test scores for mathematics but not reading. In reading, far fewer than 15 percent of all teachers perform either well or poorly, a phenomenon the testing industry has recognized well before value-added methodologies became possible, and for which there is no single accepted explanation.

A couple of explanations are likely—certainly more likely than the rather implausible assertion that reading teachers do not vary in quality as much as math teachers do. First, apart from recognizing money and being able to count it, children do not learn much math outside of school. The same cannot be said of reading, a skill that is largely dependent upon oral language skills. In fact, we learn most of the language needed to make us good readers outside of school, particularly in the first four or five years of life.

Second, standardized tests lend themselves to testing math skills much more readily than reading skills. Children learn math by algorithm and rote, processes that mirror more closely the content of tests. The process of reading and learning new words, in contrast, is far more impenetrable. Educators even debate what a reading comprehension test actually tests. One camp argues that these tests assess fluency, vocabulary, and background knowledge; another camp argues that they assess thinking skills and comprehension strategies.

A couple of recent studies that compared how teachers who did not attend a formal teacher preparation program with their traditionally prepared counterparts illustrate how a misunderstood psychometric phenomenon can have serious consequences for policy making, underscoring the need to proceed with caution when using these tests. In these studies, “alternative certification” teachers, including those from the controversial Teach For America program, produced sizable differences in the mathematics achievement of their students compared to their traditional counterparts, but in reading the two sets of teachers looked far more similar.

It may be that these alternative certification teachers were just not as good at teaching reading as their counterparts. This would explain why they couldn’t get the results in reading that they got in math. On the other hand, it may be just as likely—if not more likely given the high literacy skills of these teachers—that the reading tests themselves were the culprit. While alternative certification groups like Teach For America scramble to remedy their teachers’ flat reading scores, they may discover that finding the answer lies beyond their capacity. Reading tests are more cumulative than math tests. Students are far less likely to encounter familiar material on these tests that would reflect how well they were taught in a particular school year.

Noted education scholar E. D. Hirsch,³ long a critic of the way schools teach reading comprehension and the tests that measure this skill, contends that building the knowledge required to become a good reader requires successive, multiple years of good teaching. The impact of a single year of a great teacher, crudely measured by the ten short passages of text that typically make up the current standardized tests, may not

even be detected for several years. In other words, given the nature of language building, a reading teacher may seem to have little to no impact on students in a given year even though she may have actually made significant strides. Provided other teachers continue her good work, the results will be seen eventually, but perhaps not for several years.

Conclusion

In sum, these challenges should be cause for toning down some of the hyperbole surrounding the teacher quality findings and should instead direct our thought and energies into fair and valid reforms. For example, to use these methodologies to make teacher assignments, we need to know how to better predict who is going to be a highly effective teacher. Yet the variables that we know correlate with teacher effectiveness account for none or little of what actually makes a teacher effective. We need to be able to isolate the effect of a teacher from other influences on a classroom, all of which will take more research and study.

We also need to accept the limitations of using the current standardized tests alone as the basis for high-risk decisions like a teacher’s pay or bonus. These limitations pose legitimate fairness issues because it is simply beyond the current capacity of standardized tests or value-added methodologies to isolate the teacher’s contribution to test results. Reading tests are clumsy indicators of how much students have learned in a given year. For these tests to be used responsibly, they should be considered as one piece of evidence of a teacher’s effectiveness, part of an entire package that includes the judgments formed during frequent and well-designed evaluations by a school principal or the senior faculty.

3. Hirsch’s criticism of these tests resides in the mistaken impression that comprehension strategies can be conceptually learned and applied to any subject area, that this notion runs counter to the principles of learning supplied by cognitive psychology. What’s really being tested on a test of reading comprehension is a student’s background knowledge of a particular topic that allows him or her to read a passage with ease.

*Robbing
Peter to
Pay Paul*

6

One implication from these findings stands out from all of the others. We must improve the overall quality of the nation's teaching force. If schooling can move beyond its theoretical impact to actually overcome a child's family environment, we must increase the number of great teachers. If one out of every five teachers met this standard instead of one out of every seven, a child's chance of being randomly assigned a great teacher for five years in a row would decrease from 1 in 17,000 to 1 in 3,125. That number is still high odds, but it makes it much more practical to make conscientious policy decisions that assign the best teachers to the children that need them the most. A typical faculty of 35 teachers would have seven instead of five phenomenal teachers, enough for each grade.

To increase the number of great teachers requires that we broaden the current sources of new teachers. Schools of education fall short both in the quantity and quality of the teachers they prepare because they are attracting fewer and fewer teacher candidates who are themselves academically able. State policies for teacher preparation and licensure need to be overhauled substantially. The National Council of Teacher Quality will be releasing an annual report card on state teacher policies beginning in September 2006, describing in concrete, practical terms how policies need to change. Collective bargaining agreements that support a uniform salary schedule and sloppy hiring practices in districts (which permit interview protocols that border on negligence) must also change. Only when these problems are fixed will the profession be able to attract, recruit, and keep good teachers.