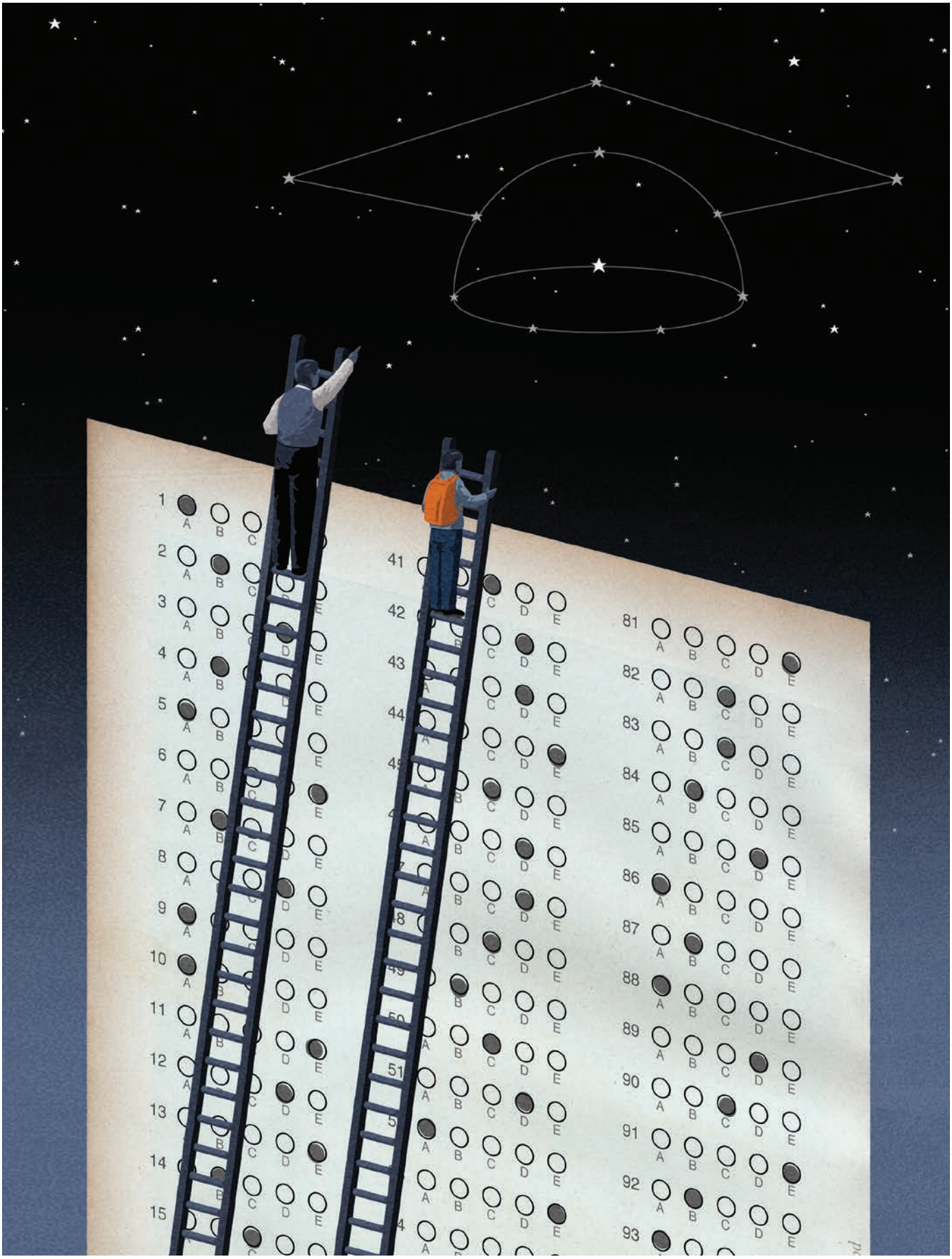


ILLUSTRATION / MICHAEL GLENWOOD

THE FULL MEASURE OF A TEACHER

USING VALUE-ADDED to assess effects on student behavior

WHEN STUDENTS LOOK BACK on their most important teachers, the social aspects of their education are often what they recall. Learning to set goals, take risks and responsibility, or simply believe in oneself are often fodder for fond thanks—alongside mastering precalculus, becoming a critical reader, or remembering the capital of Turkmenistan.

It's a dynamic mix, one that captures the broad charge of a teacher: to teach students the skills they'll need to be productive adults. But what, exactly, are these skills? And how can we determine which teachers are most effective in building them?

Test scores are often the best available measure of student progress, but they do not capture every skill needed in adulthood. A growing research base shows that non-cognitive (or socio-emotional) skills like adaptability, motivation, and self-restraint are key determinants of adult outcomes. Therefore, if we want to identify good teachers, we ought to look at how teachers affect their students' development across a range of skills—both academic and non-cognitive.

A robust data set on 9th-grade students in North Carolina allows me to do just that. First, I create a measure of non-cognitive skills based on students' behavior in high school, such as suspensions and on-time grade progression. I then calculate effectiveness ratings based on teachers' impacts on both test scores and non-cognitive skills and look for connections between the two. Finally, I explore the extent to which measuring teacher impacts on behavior allows us to better identify those truly excellent educators who have long-lasting effects on their students.

I find that, while teachers have notable effects on both test scores and non-cognitive skills, their impact

by C. KIRABO JACKSON

on non-cognitive skills is 10 times more predictive of students' longer-term success in high school than their impact on test scores. We cannot identify the teachers who matter most by using test-score impacts alone, because many teachers who raise test scores do not improve non-cognitive skills, and vice versa.

These results provide hard evidence that measuring teachers' impact through their students' test scores captures only a fraction of their overall effect on student success. To fully assess teacher performance, policymakers should consider measures of a broad range of student skills, classroom observations, and responsiveness to feedback alongside effectiveness ratings based on test scores.

A Broad Notion of Teacher Effectiveness

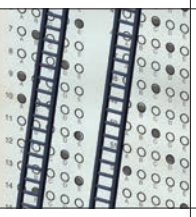
Individual teacher effectiveness has become a major focus of school-improvement efforts over the last decade, driven in part by research showing that teachers who boost students' test scores also affect their success as adults, including being more likely to go to college, have a job, and save for retirement (see "Great Teaching," *research,* Summer 2012). Economists and policymakers have used students' standardized test scores to develop measures of teacher performance, chiefly through a formula called value-added. Value-added models calculate individual teachers' impacts on student learning by charting student progress against what they would ordinarily be expected to achieve, controlling for a host of factors. Teachers whose students consistently beat those odds are considered to have high value-added, while those whose students consistently don't do as well as expected have low value-added.

At the same time, policymakers and educators are focused on the importance of student skills not captured by standardized tests, such as perseverance and collaborating with others, for longer-term adult outcomes. The 2015 federal Every Student Succeeds Act allows states to consider how well schools do at helping students create "learning mindsets," or the non-cognitive skills and habits that are associated with positive outcomes in adulthood. In one major experiment in California, for example, a group of large districts is tracking progress in students' non-cognitive skills as part of their reform efforts.

Is it possible to combine these two ideas by determining which individual teachers are most effective at helping students develop non-cognitive skills?

To examine this question, I look to North Carolina, which collects data on test scores and a range of student behavior. I use data on all public-school 9th-grade students between 2005 and 2012, including demographics, transcript data, test scores in grades 7 through 9, and codes linking scores to the teacher who administered the test. The data cover about 574,000 students in 872 high schools. I focus on the 93 percent of 9th-grade students who took classes in which teachers will also have traditional test score-based value-added ratings: English I and one of three math classes (algebra I, geometry, or algebra II).

I use these data to explore three major questions. First, how predictive is student behavior in 9th grade of later success in high school, compared to student test scores? Second, are teachers who are better at raising test scores also better at improving student behavior? And finally, what measure of teacher performance is more predictive of students' long-term success: impacts on test scores, or impacts on non-cognitive skills?



IF WE WANT TO IDENTIFY GOOD TEACHERS,
we ought to look at how teachers affect their students' development across a range of skills—both academic and non-cognitive.

The Predictive Power of Student Behavior

To explore the first question, I create a measure of students' non-cognitive skills by using the information on their behavior available in the 9th-grade data, including the number of absences and suspensions, grade point average, and on-time progression to 10th grade. I refer to this weighted average as the "behavior index." The basic logic of this approach is as follows: in the same way that one infers that a student who scores higher on tests likely has higher cognitive skills than a student who does not, one can infer that a student who acts out, skips class, and fails to hand in homework *likely* has lower non-cognitive skills than a student who does not. I also create a test-score index that is the average of 9th-grade math and English scores.

I then look at how both test scores and the behavior index are related to various measures of high-school success, using administrative data that follow students' trajectories over time. The outcomes I consider include

graduating high school on time, grade-point average at graduation, taking the SAT, and reported intentions to enroll in a four-year college. Roughly 82 percent of students graduated, 4 percent are recorded as having dropped out, and the rest either moved out of state or remained in school beyond their expected graduation year. Because I am interested in how changes in these skill measures predict long-run outcomes, I control for the student's test scores and behavior in 8th grade. In addition, my analysis adjusts for differences in parental education, gender, and race/ethnicity.

My first set of results shows that a student's behavior index is a much stronger predictor of future success than her test scores. Figure 1 plots the extent to which increasing test scores and the behavior index by one standard deviation, equivalent to moving a student's score from the median to the 85th percentile on each measure, predicts improvements in various outcomes. A student whose 9th-grade behavior index is at the 85th percentile is a sizable 15.8 percentage points more likely to graduate from high school on time than a student with a median behavior index score. I find a weaker relationship with test scores: a student at the 85th percentile is only 1.9 percentage points more likely to graduate from high school than a student whose score is at the median. The behavior index is also a better predictor than 9th-grade test scores of high-school GPA and the likelihood that a student takes the SAT and plans to attend college.

While these patterns reveal that the behavior index is a good predictor of educational attainment, they are descriptive. They do not show that teachers impact these behavior, and they do not show that teacher impacts on these measures will translate into improved longer-run success. I next examine these more causal questions.

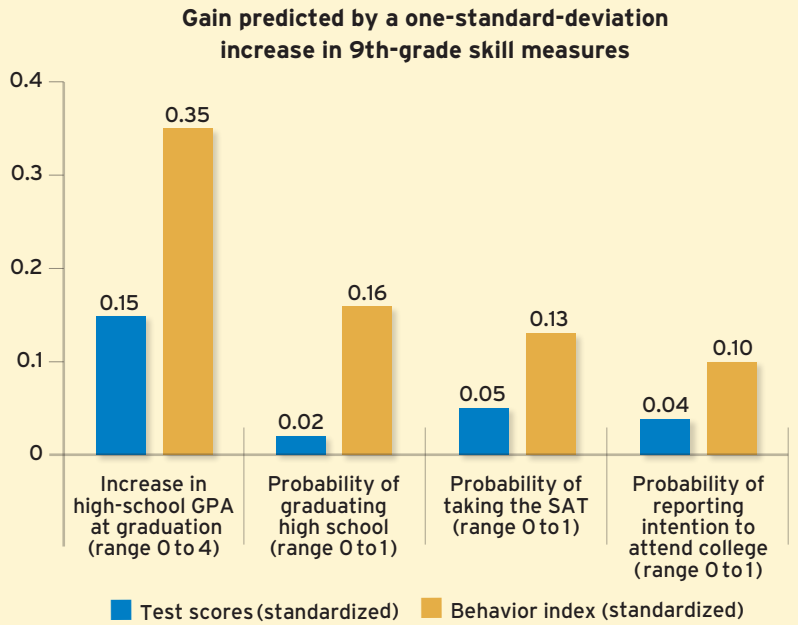
Applying Value-Added to Non-Cognitive Skills

The predictive power of the behavior index suggests that improving behavior could yield large benefits, but it leaves open the question of whether teachers who improve student behavior are different from teachers who improve test scores. This is important, because if teachers who are more effective at raising test scores are also more effective at improving behavior, then we will not improve our ability to identify teachers who improve long-run student outcomes by estimating teacher impacts on behavior. In contrast, if the group of teachers who are effective at improving test scores includes some who are above average, average, or even below average at improving behavior, then having non-cognitive effectiveness ratings will allow us to identify truly excellent teachers who may have the largest impact on longer-run outcomes by improving *both* test scores and behavior.

To assess this, I employ separate value-added models to evaluate the unique contribution of individual teachers to test scores and to the behavior index. I group teachers by their ability to improve behavior, and plot the distribution of test-score value-added among teachers in each group. If teachers who improve one skill are

Behavior Is a Stronger Predictor of Student Success (Figure 1)

Improving a 9th-grade student's rank on the behavior index predicts greater high-school and college-going success than moving a student's test scores the same amount.

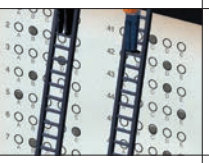


NOTE: Analysis adjusts for differences in students' test scores and behavior in 8th grade, gender, race/ethnicity, and parental education.

SOURCE: Author's calculations using information from the North Carolina Education Research Data Center

also those who improve the other, the average test-score value-added should be much higher in groups with higher behavior value-added, and there should be little overlap in the distribution of test-score value-added across the behavior value-added groups.

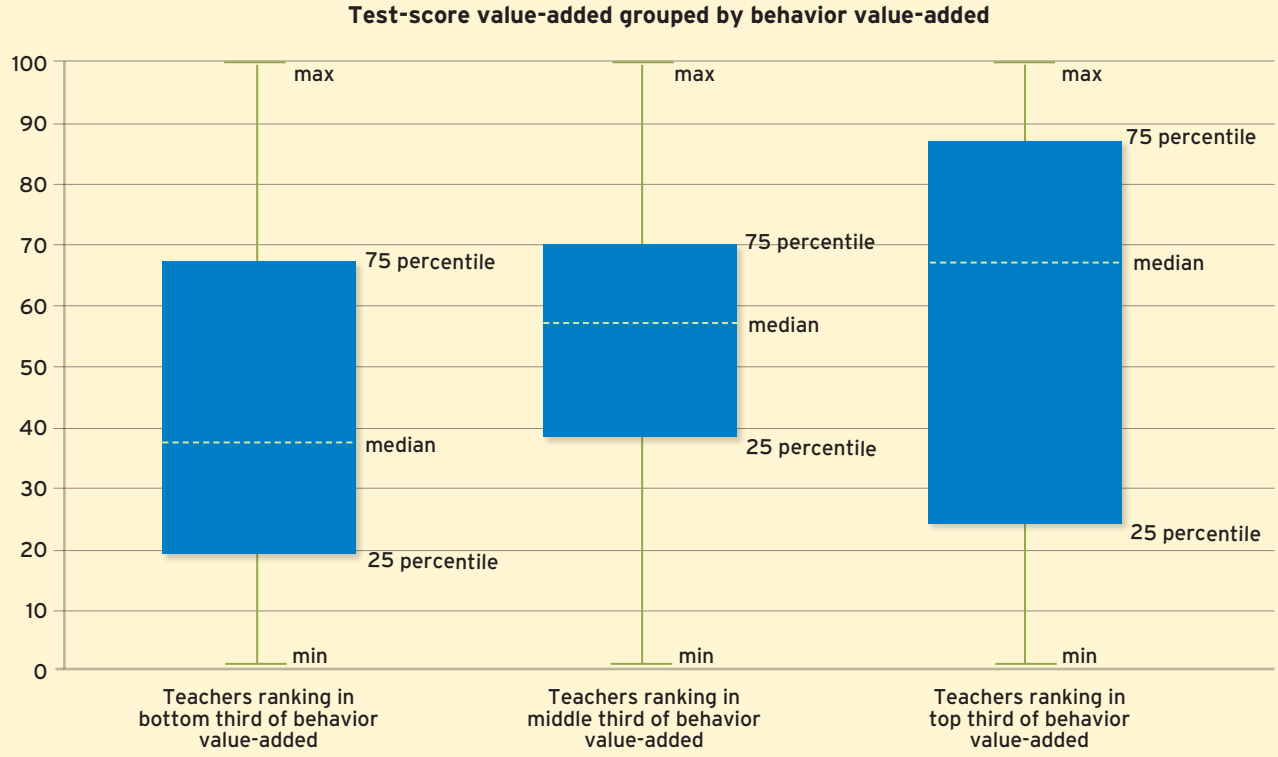
That's not what the data show. Although teachers with higher behavior value-added tend to have somewhat higher test-score value-added, there is considerable overlap across groups (see Figure 2). That is, although teachers who are better at raising test scores tend to be better at raising the behavior index, *on average,* effectiveness along one dimension is a poor predictor of the other. For example, among the bottom third of teachers with the worst behavior value-added, nearly 40 percent are above average in test-score value-added. Similarly, among the top third of teachers with the best behavior value-added, only 58 percent of teachers are above average in test-score value-added. This reveals not only that many teachers who are excellent at improving one skill are poor at improving the other, but also that knowing a teacher's impact on one skill provides little information on the teacher's impact on the other.



A STUDENT'S BEHAVIOR INDEX IS A MUCH STRONGER PREDICTOR of future success than her test scores.

Many Teachers Who Excel at Improving Test Scores Do Not Improve Behavior (Figure 2)

Only 58 percent of teachers in the top third in improving student behavior rank above average in test-score value-added.



SOURCE: Author's calculations using information from the North Carolina Education Research Data Center

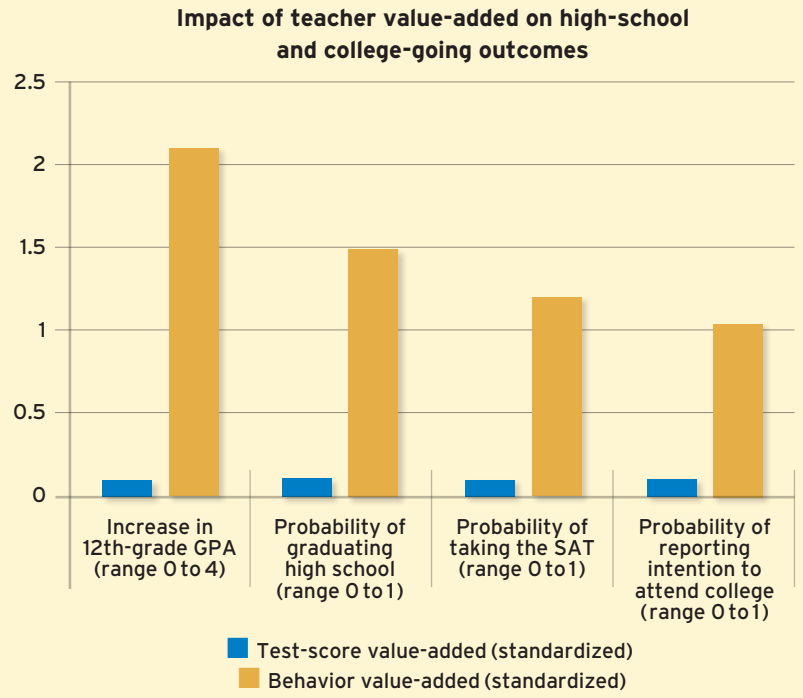
Impacts on High-School Success

The patterns I have documented so far suggest that there could be considerable gains to using teacher impacts on both test scores and behavior to identify teachers who may improve longer-run outcomes. To assess this directly, I examine the extent to which the estimated value-added of a student's teacher in 9th grade influenced his or her outcomes at the end of high school, such as graduating on time, taking the SAT, and planning to go to college. To avoid biases, I use a teacher's value-added based on her impacts *in other years* as my measure of teacher effectiveness. I then estimate two impacts on students' longer-run outcomes: the impact of having a teacher whose test-score value-added is one standard deviation higher than the median, and that of having a teacher whose behavior value-added is one standard deviation higher.

A teacher's value-added to 9th-grade behavior is a much stronger predictor of her impacts on subsequent educational attainment than her impacts on 9th-grade test scores (see Figure 3). For example, having a teacher at the 85th percentile of test-score value-added would increase a student's chances of graduating high school on time by about 0.12 percentage points compared to having an average teacher. In contrast, having a teacher at the 85th percentile of behavior value-added would increase high-school graduation by about 1.46 percentage points compared to having an average teacher.

Teachers Who Improve Student Behavior Matter Most (Figure 3)

A teacher's impact on behavior is about 10 times more predictive of her impact on high-school graduation than her impact on test scores. This pattern holds true across multiple measures of high-school success.



NOTE: Figure shows the predicted effects of having a teacher who is one standard deviation more effective in improving test scores to the behavior index.

SOURCE: Author's calculations using information from the North Carolina Education Research Data Center

In other words, the impact of teachers on behavior is about 10 times more predictive of whether they increase students' high-school completion than their impacts on test scores. This basic pattern holds true for all of the longer-run outcomes examined, including plans to attend college. Remarkably, the causal estimates in Figure 3 are almost exactly what one might have expected from the descriptive patterns in Figure 1.

These results confirm an idea that many believe to be true but that has not been previously documented—that teacher effects on test scores capture only a fraction of their impact on their students. The fact that teacher impacts on behavior are much stronger predictors of their impact on longer-run outcomes than test-score impacts, and that teacher impacts on test scores and those on behavior are largely unrelated, means that the lion's share of truly excellent teachers—those who improve long-run outcomes—will not be identified using test-score value-added alone.

To make this point more concretely, I look at another group of teachers: those in the top 10 percent based on their impacts on high-school graduation. I then look at whether these teachers are also in the top 10 percent based on their test-score value-added and their behavior value-added. Behavior value-added does a much better job of identifying those teachers that improve on-time graduation: 93 percent of teachers in the top 10 percent with respect to graduation are also in the top 10 percent of behavior

value-added. Only 20 percent of these high-impact teachers are in the top 10 percent of test-score value-added.

At the other end of the performance spectrum, behavior value-added also is better at identifying teachers who are the worst at improving students' chances of graduating high school on time. Among the bottom 10 percent of teachers with the lowest predicted impacts on high-school graduation, 89 percent are in the bottom 10 percent of behavior value-added and 32 percent are in the bottom 10 percent of test-score value-added.

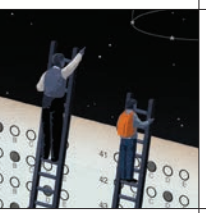
Implications

Teachers are more than educational-outcome machines—they are leaders who can guide students toward a purposeful adulthood. This analysis provides the first hard evidence that such contributions to student progress are both measurable and consequential. That is not to say that test scores should not be used in evaluating teacher effectiveness. Test-score impacts are an important gauge of teacher effectiveness for the roughly one in five teachers who teach grade levels and subjects in which it is possible to construct test-based value-added ratings. Impacts on behavior are another critical measure for those teachers and for everyone else, and can serve as an additional source of information in a strong multiple-measures evaluation system, which could also include observations, student surveys, and evidence of responsiveness to feedback.

Using value-added modeling in this way can yield critical and novel information about teacher performance. Although the teacher characteristics available in the North Carolina data—years of teaching experience, full certification, teaching exams scores, regular licensing, and college selectivity (as measured by the 75th percentile of the SAT scores at the teacher's college)—do predict effects on test scores, none are significantly related to effects on behavior. However, this does not preclude the use of more detailed information on teachers to better predict effects on a broad range of skills; with more research, schools and districts might learn which characteristics to look for in order to hire and nurture teachers who are likely to improve students' non-cognitive skills.

Another potential application of this approach to measuring teacher effectiveness could be to provide incentives that districts could offer teachers to improve student behavior. However, because some of the behavior can be "improved" by changes in teachers' practices that do not improve student skills, such as inflating grades and misreporting misconduct, attaching external stakes to measures of students' non-cognitive skills may not be beneficial—at least not without addressing this "gameability" problem.

One possibility is to find measures of non-cognitive skills that are difficult to adjust unethically. For example, classroom observations and student and parent surveys may provide valuable information about student skills that are not measured by test scores and are less easily manipulated by teachers. Policymakers could attach incentives to both these measures of non-cognitive skills and test scores to promote better longer-run outcomes. Another approach is to provide teachers with incentives to improve the behavior of students in their classrooms the *following* year, when the teacher's influence may still be present but the teacher can no longer manipulate the measurement of student behavior. Or, policymakers could identify teaching practices that improve behavior and provide incentives for teachers to engage in these practices. Such approaches have been used successfully to increase test scores (see "Can Teacher Evaluation Improve Teaching?" *research,* Fall 2012).



ALTHOUGH TEACHERS WHO ARE BETTER AT RAISING TEST SCORES tend to be better at raising the behavior index, on average, effectiveness along one dimension is a poor predictor of the other.

Teachers influence just the sort of non-cognitive skills that research shows boost students' success through high school and beyond. And through value-added modeling, we can estimate individual impacts and unearth another piece of the teacher-performance puzzle. Although the policy path ahead is not immediately clear, the fact that teachers have impacts on a set of skills that predict longer-run success but are not captured by current evaluation methods is important. The findings suggest that any policy to identify effective teachers—whether for evaluation, targeted professional development, or de-selection—should seek to use teacher impacts on a broader set of outcomes than test scores alone.

C. Kirabo Jackson is professor of human development and social policy at Northwestern University. This article is based on "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes," Journal of Political Economy, *2018, Vol. 126, No. 5.*