

Are Student Surveys the Right Tools for Evaluating Teacher Performance?

Yes. No. Maybe.

IN THE 1990s, standardized tests became entrenched in American K–12 schools as nearly every state, and later the federal government, adopted policies that mandated annual testing and held schools accountable for the results. In the ensuing decades, however, educators and policymakers began to recognize that high-stakes testing was not living up to its promise and that the single-minded focus on test scores had produced unintended (although, in retrospect, entirely predictable) consequences.

Increasingly, school districts across the country are now turning to an alternative evaluation tool—surveys that ask students to rate their teachers and their schools on various metrics of quality and effectiveness. This growing use of evaluative surveys in K–12 reflects a rare consensus among education policy wonks and activists, bringing together strange ideological bedfellows who all believe surveys can help achieve their goals and priorities.

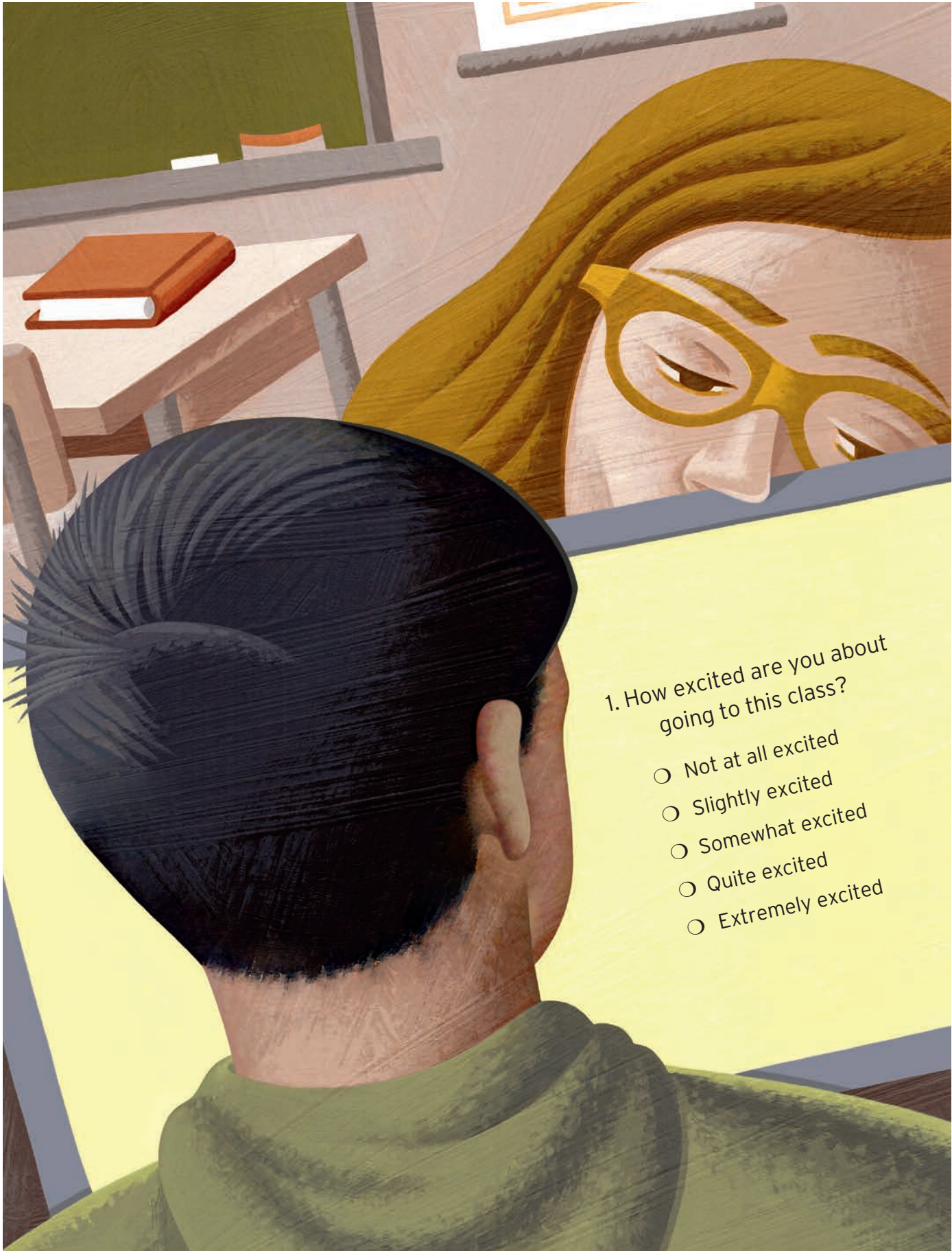
Unfortunately, there is a risk that education leaders will make the same mistakes with surveys that they did with standardized tests—overpromising and not thinking through perverse incentives. Fortunately, it's not too late to consider carefully both the promise and the likely pitfalls of using student surveys as a measure of teacher and school performance.

Judging Teachers

Education research has established that teachers are the most important in-school factor influencing student academic achievement. The same research, however, documents considerable variation in the effectiveness of public school teachers, suggesting that improving the workforce—by providing professional development for existing educators, recruiting better teachers through nontraditional pathways, and dismissing the poorest performers—offers a promising policy lever for raising student outcomes. Many states reformed their teacher-evaluation policies during the 2010s, after the Obama administration launched its Race to the Top grant competition, which incentivized states to adopt rigorous evaluation systems designed to measure and reward teacher contributions to student learning.

This effort did not work out as hoped. With a few notable exceptions, such as the highly regarded IMPACT system in Washington, D.C., it seems that efforts to improve teacher rating systems have largely been a bust. One recent analysis of state-level teacher-evaluation reforms found “precisely estimated null effects.” Commentators have offered many hypotheses as to why these initiatives fell short, but one probable explanation is that the metric of teacher quality preferred by

By **VLADIMIR KOGAN**



1. How excited are you about going to this class?
- Not at all excited
 - Slightly excited
 - Somewhat excited
 - Quite excited
 - Extremely excited

reformers—“value added” to student test scores—can only be calculated for a minority of teachers, since most do not teach grade levels and subjects where standardized tests are administered annually. The ensuing push for one-size-fits-all evaluation systems resulted in considerable weight being put on other, more easily gameable or subjective measures of performance that could be applied to more teachers.

That is one reason why some accountability hawks are now pinning their hopes on student surveys, which can be administered in every subject and to students as young as grade 3. The innovative teacher evaluation system in Dallas, identified as one contributor to recent improvements recorded by the city’s lowest-performing schools and described as a national model by some reformers, relies heavily on student surveys. The Dallas survey of students in grades 6–12 asks them to evaluate factors such as the teacher’s expectations of students, the positive or negative “energy” in the classroom, the fairness of the teacher’s rules, the depth of a teacher’s

subject knowledge, the frequency of helpful feedback, the clarity of instruction, and more.

Critics of standardized testing have also written favorably about student surveys, arguing that they help move education leaders beyond the obsessive focus on test scores by identifying other aspects of teacher and school quality valued by students, parents, and policymakers. One of the most influential researchers in this area is Northwestern University economist Kirabo Jackson (an *Education Next* contributor). In pathbreaking work, Jackson showed that measures of teacher quality based narrowly on contributions to test-score improvement missed many other ways teachers affect long-run student outcomes. More recently, Jackson used data from Chicago high schools to show that student surveys can help quantify important dimensions of school quality, including school climate, that affect not just student achievement but also outcomes such as high school graduation rates and criminal-justice involvement. Jackson’s recent appointment to President

PANORAMA
EDUCATION

DALLAS
INDEPENDENT SCHOOL DISTRICT

Student Survey - Teacher

For the following questions, we are interested in learning more about how you think about your teachers and your classes. Please give us your honest opinions. With your thoughtful feedback we can make your school a better place for learning.

Perceptions of Class
Please give us your opinions on this class so that we can better understand your experience in this class.

1. How excited are you about going to this class?

Not at all excited Slightly excited Somewhat excited Quite excited Extremely excited

2. How positive or negative is the energy of this class?

Very negative Somewhat negative Slightly negative Neither negative nor positive Slightly positive Somewhat positive Very positive

3. How often does this teacher take time to make sure you understand the material?

Almost never Once in a while Sometimes Frequently Almost always

The Student Experience Survey for students in grades 6 to 12 in the Dallas Independent School District asks them how they feel about their class and the teacher. Such teacher evaluation systems are credited with helping to improve the city’s lowest-performing schools.

The innovative teacher evaluation system in Dallas, identified as one contributor to recent improvements recorded by the city’s lowest-performing schools, relies heavily on student surveys.

Biden’s Council of Economic Advisors suggests that survey-based measures are likely to play a bigger role in federal school-improvement efforts in the future.

Student surveys also play a central role in policies promoted by many other political entrepreneurs. For example, on the political left, increasing interest in social and emotional learning will also mean greater reliance on student surveys, since they represent one of the few ways in which such skills can be measured and quantified. At the same time, conservatives have embraced surveys in their efforts to promote free speech and protect ideological diversity in schools. Proposed legislation in Ohio, based on model bills developed by high-profile conservative think tanks, would require that public university professors have their teaching evaluated in large part through student surveys, including a specific question asking, “Does the faculty member create a classroom atmosphere free of political, racial, gender, and religious bias?”

Too Much Too Fast?

Promising as these developments may seem, it is concerning that the hype surrounding student surveys has gotten well ahead of the evidence. Researchers have devoted too little attention to validating survey-based measurements to confirm that they assess the things policymakers hope to measure. Nor have decisionmakers sufficiently considered the potential consequences of attaching high stakes to student survey responses. (Jackson’s work in Chicago sheds little light on this question, as it was conducted at a time when surveys were not part of the city’s school accountability system.)

One cautionary piece of evidence comes from the Gates Foundation–funded Measures

of Effective Teaching project. As part of this effort, researchers compared three distinct ways of assessing teacher quality—test-score value-added, classroom observations, and student surveys. While early data did find some evidence that survey-based measures predicted test-score growth, these results were not confirmed in the more rigorous part of the study in which students were randomly assigned to different teachers. The final results found no relationship between student survey scores and improvements in academic achievement, prompting researchers to suggest “practitioners should proceed with caution when considering student survey measures for teacher evaluation.”

Other potential problems also need scrutiny. For example, one recent study examined the association of survey-based measures of student conscientiousness, self-control, and grit with outcomes such as school attendance, disciplinary infractions, and gains in test scores over time. While researchers found a positive relationship between attitudes and behavioral outcomes among students attending the same schools, these correlations disappeared when the same data were aggregated up to the school level and compared across campuses. Most worrying, the authors also found that high-performing charter schools, shown through randomized lotteries to improve both student attendance



NORTHWESTERN UNIVERSITY

Kirabo Jackson’s research showed that student surveys helped quantify how schools affected graduation rates and subsequent criminal justice involvement.

Researchers have devoted too little attention to validating survey-based measurements to confirm that they assess the things policymakers hope to measure.

and academic achievement, recorded the *lowest* scores on the student surveys. One possible explanation is that the school environment may have affected survey responses in unexpected ways—with students in classes made up of higher-performing peers rating their own attributes more critically, through a form of negative social comparison.

Such results are unlikely to surprise political pollsters, who have long understood the importance of both priming and framing effects in shaping survey responses. That is, even modest changes in the survey-taking context—such as changing the order of the questions—can have a significant impact on the responses. Designing survey questions that actually measure what their authors intend to measure requires considerable skill. Small variations in question wording—for example, describing a protest as an exercise in free speech as opposed to a threat to public safety—can yield sharply different results. Unfortunately, too few education practitioners working with student survey data have any rigorous training in survey research methods.

Finally, although many now appreciate the ways in which high-stakes accountability policies can encourage “teaching to the test,” few have considered the problem of “teaching to the survey.” Letting students weigh in on teacher evaluations, as is done under the Dallas model, is a great way to encourage teachers to do more of what students want. But whether those changes lead to improvements in instructional quality is another matter, and there are many reasons to expect that they won’t.

Lessons from Other Fields

Fields outside of primary and secondary education that have used evaluative surveys for decades provide disturbing examples of undesirable and problematic gaming behaviors that such surveys can incentivize. At the college level, student evaluations have long served as the primary method for evaluating teaching, and considerable evidence indicates that this practice has contributed to grade inflation. Regardless of the specific questions included in the survey, student responses appear to reflect

their satisfaction with grades (higher is better!) and the effort required in the course (less is better!). Some professors have even resorted to bringing sweets to class on days when students complete their surveys, as such treats seem to significantly boost evaluation scores.

As Doug Lemov has argued, grading reforms implemented during the pandemic in hopes of reducing stress and supporting teenage mental health have contributed to grade compression and diluted the returns to student effort (see “Your Neighborhood School Is a National Security Risk,” *features*, Winter 2024). The experience from higher education suggests that incorporating student surveys into formal teacher evaluations will only exacerbate these dynamics.

Although some equity advocates have reacted with alarm to recent research finding racial gaps in principals’ evaluations of teachers, systemic bias—against women, nonwhite professors, and nonnative English speakers—has long been documented in student-survey evaluations of college instructors. Ironically, growing interest in inherently subjective surveys coincides with technological changes, including using AI to classify and score recorded lesson videos, that promise to remove much of the personal discretion from teaching observations.

Even more concerning evidence comes from the field of medicine, where patient satisfaction surveys are required for hospital accreditation and, since the passage of the Affordable Care Act, linked to Medicare reimbursements. For example, some studies suggest that patients rate doctors more favorably when they prescribe antibiotics on demand, including for viral colds for which this treatment is inappropriate because it may contribute to the rise of antibiotic resistance in the population. One journalist has argued that, because a number of the patient-satisfaction questions ask about pain management, the use of high-stakes surveys has also contributed to America’s opioid epidemic by creating pressure on doctors to overprescribe pain pills in order to achieve higher ratings.

If there is one lesson that the past four decades of education reform have taught us, it’s that well-meaning policies rarely work as their proponents expect and hope. Sometimes they even backfire, producing the opposite of what was intended. Both practitioners and policymakers should remember these lessons as they think about how to incorporate student surveys into education-accountability systems or use such data to shape policy.

Vladimir Kogan is a professor in The Ohio State University’s Department of Political Science and (by courtesy) the John Glenn College of Public Affairs.