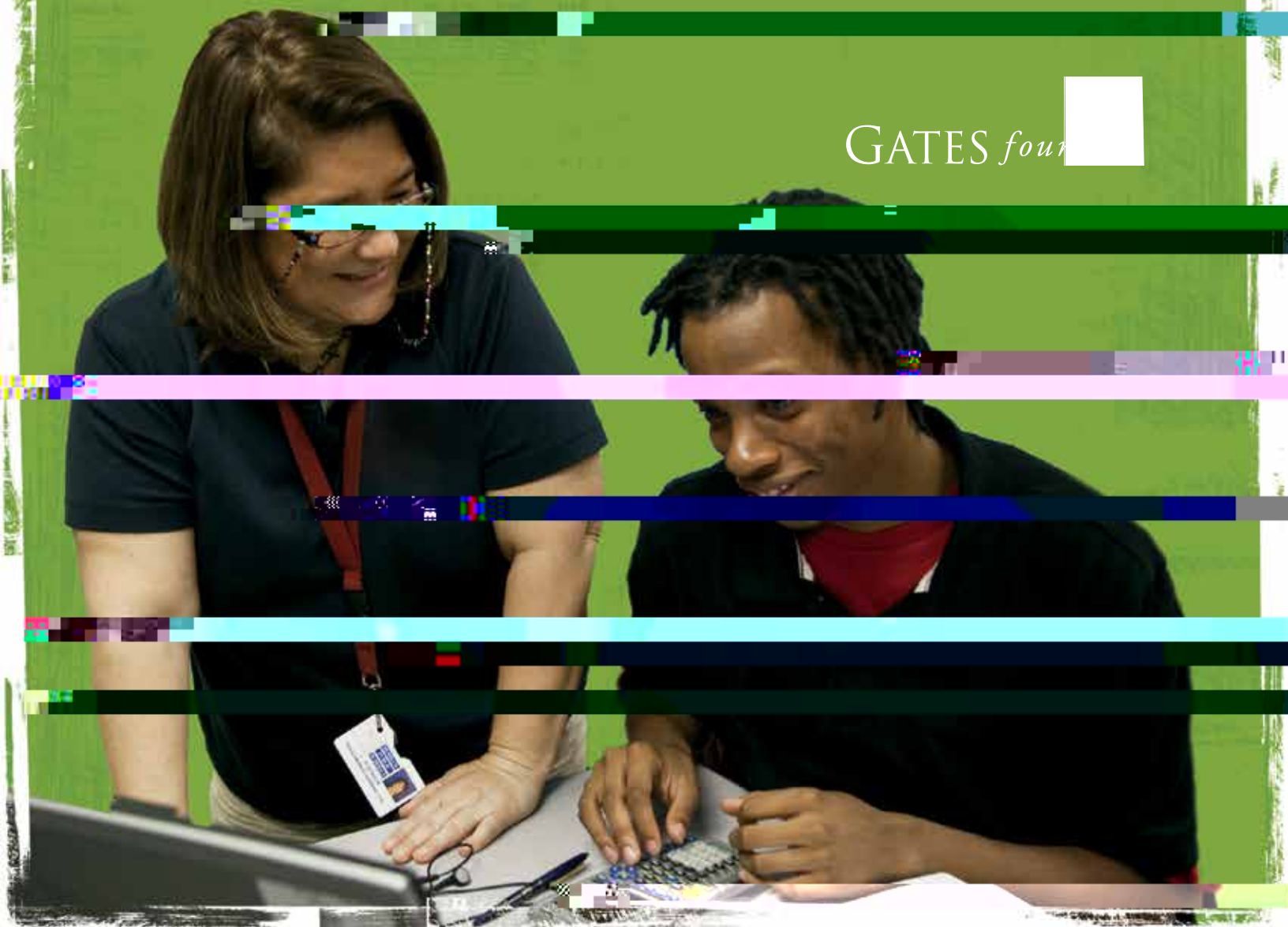
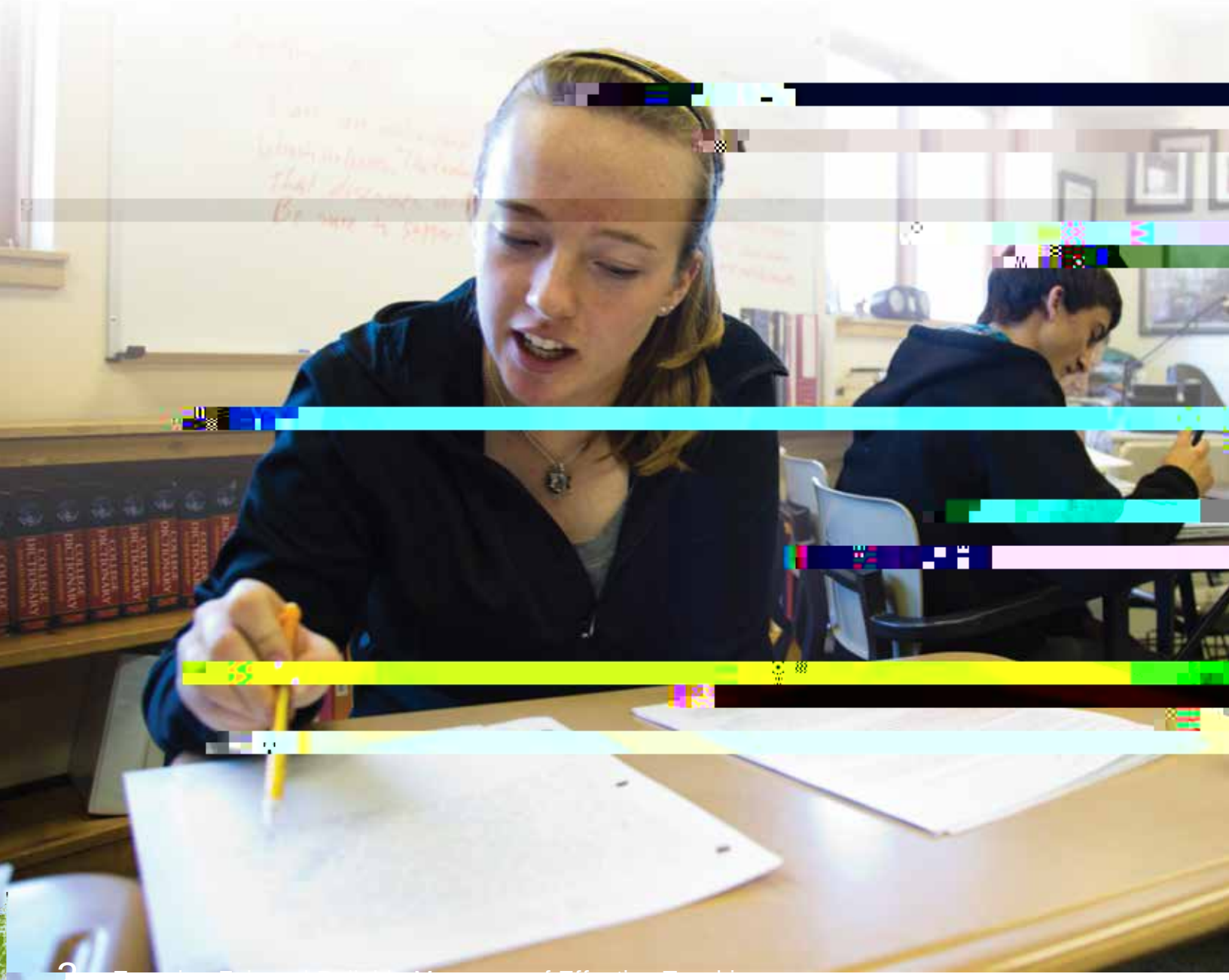
MET
project

GATES foun



ABOUT THIS REPORT:

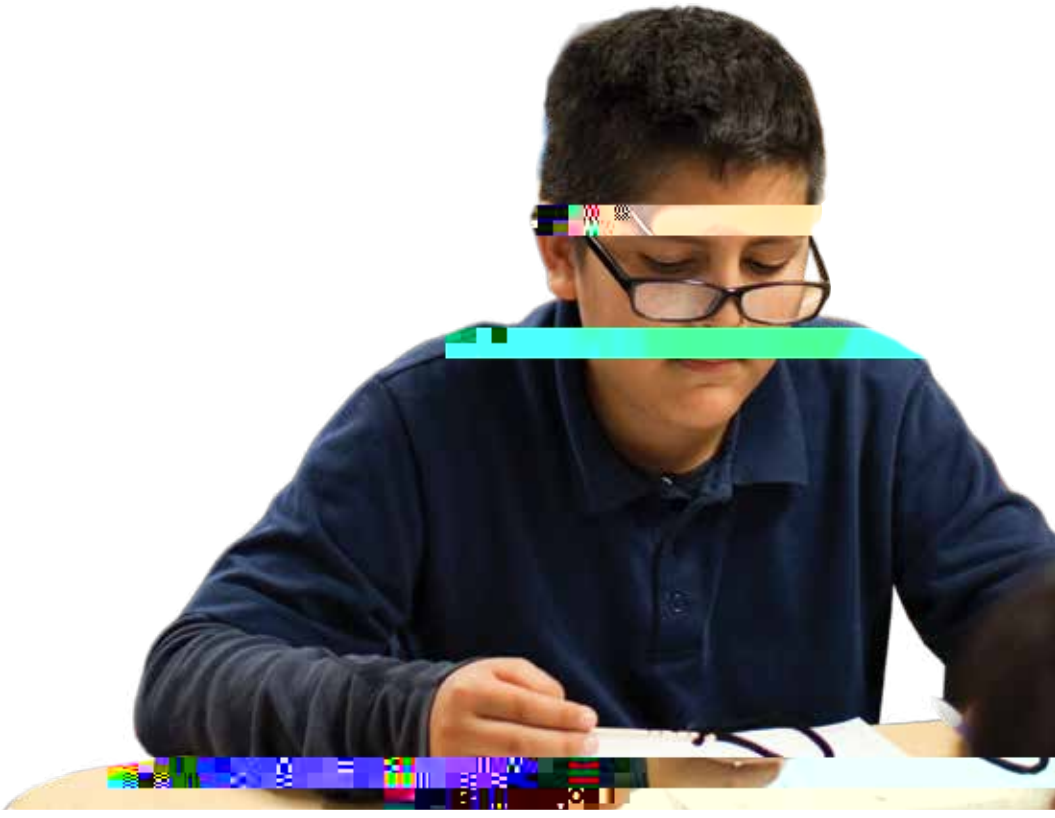
Contents



States and districts have launched unprecedented efforts in recent years to

Can measures of effective teaching identify teachers who better help students learn?

Despite decades of research suggesting that teachers are the most important in-school factor affecting student learning,



By definition, teaching is effective when it enables student learning. But identifying effective teaching is complicated by the fact that teachers often have very different students. Students start the year with different achievement levels and different needs. Moreover, some teachers tend to get particular types of students year after year (that is, they tend to get higher-performing or lower-performing ones). This is why so-called value-added measures attempt to account for differences in the measurable characteristics of a teacher's students, such as prior test scouch a,

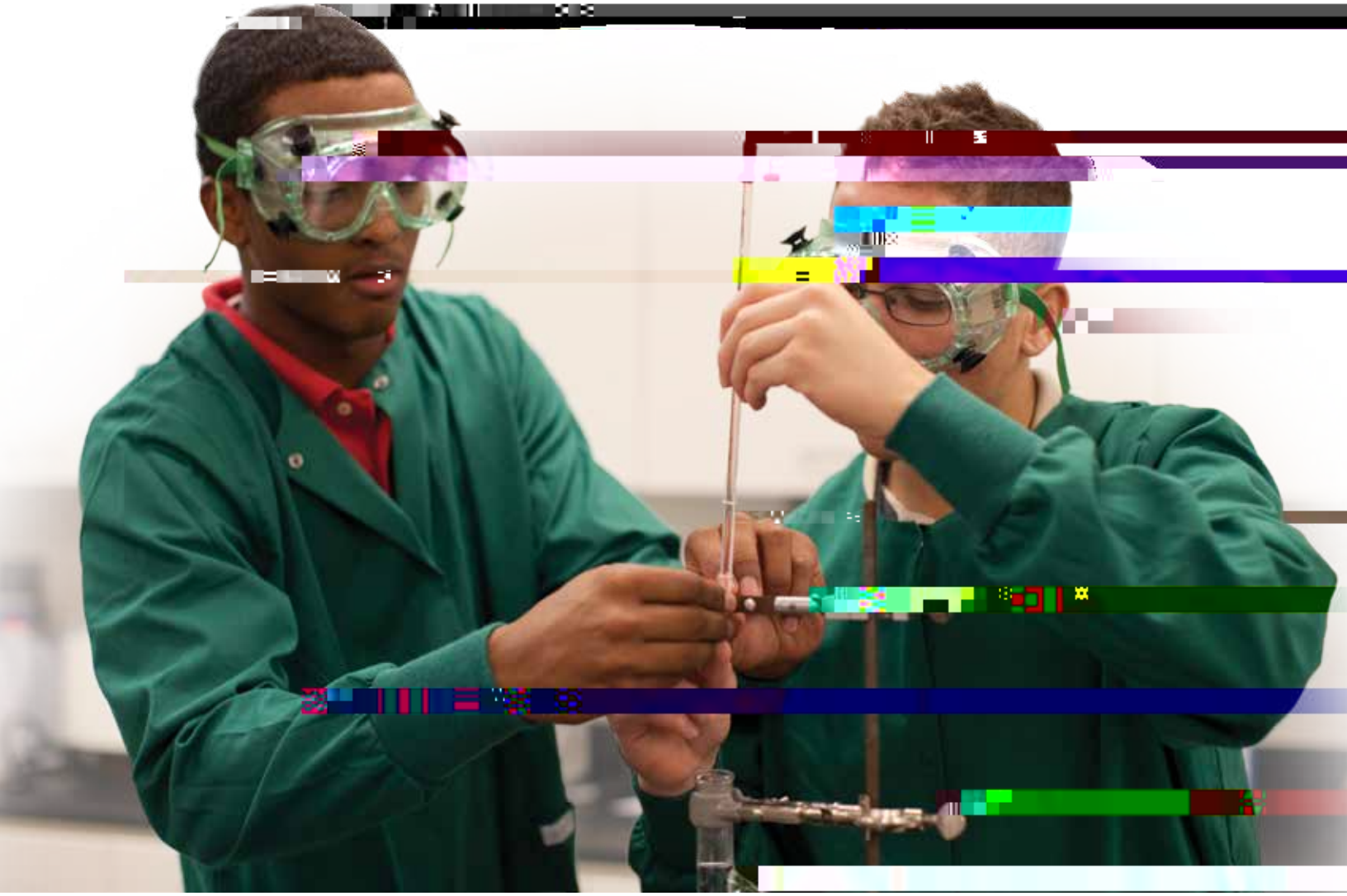
represents 5 percent of the teachers in the analysis, sorted based on their pre-

> "We can unambiguously say that school systems should adjust their achievement gain measures to account for the prior test scores of students. When we removed this control, we wound up predicting much larger differences in achievement than

controls, we cannot determine from our evidence whether school systems should include them. Our results were ambiguous on that score.

To avoid over-interpretation of these results, we hasten to add two caveats: First, a prediction can be correct on average but still be subject to measurement error. Our predictions of students' achievement following random assignment were correct on average, but within every group there were some teachers whose students performed better than predicted and some whose students performed worse. Second, we could not, as a practical matter, randomly assign students or teachers to a different school site. As a result, our study does not allow us to investigate bias in teacher effectiveness measures arising from student sorting between different schools.[6]

Nonetheless, our analysis should give heart to those who have invested considerable effort to develop practices and policies to measure and support effective teaching. Through this large-scale study involving random assignment of teachers to students, we are con dent that we can identify groups of teachers who are comparatively more effective than their peers in helping students learn. Great teaching does make a difference.



How Much Weight Should Be Place on Each Measure of E ective Teachi

Teaching is too complex for any single measure of performance to capture it accurately. Identifying great teachers requires multiple measures. While states and districts embrace multiple measures for targeted feedback, many also are combining measures into a single index, or composite. An index or composite can be a useful summary of complex information to support decisionmaking. The challenge is to combine measures in ways that support effective teaching while avoiding such unintended consequences as too-narrow a focus on one aspect of effective teaching.

To date, there has been little empirical evidence to suggest a rationale for particular weights. The MET project's report Gathering Feedback for Teaching showed that equally weighting three measures, including achievement gains, did a better job predicting teachers' success (across several student outcomes) than teachers' years of experience and masters' degrees. But that work did not attempt to determine optimal weights for composite measures.

Over the past year, a team of MET project researchers from the RAND Corporation and Dartmouth College used MET project data to compare differently weighted composites and study the implications of different weighting schemes for different outcomes. As in the Gathering Feedback for Teaching report, these composites included student achievement gains based on state assessments, classroom observations, and student surveys. The researchers estimated the ability of variously weighted composites to produce consistent results and accurately forecast

of effective teaching and neglect its other important aspects. For example, a singular focus on state tests could displace gains on other harder-to-measure outcomes. Moreover, if the goal is for students to meet a broader set of learning objectives than are measured by a state's tests, then too-heavily weighting that test could make it harder to identify teachers who are producing other valued outcomes.

Composites Compared

The research team compared four different weighting models, illustrated in **Figure 3:** (Model 1) The "best

predicting teachers' student achievement gains on state tests. By de nition, the best composite in this regard is Model 1, the model weighted for maximizing accuracy on state test results. Models 2–4 show the effect of reducing weights on student achievement gains on state tests for middle school ELA. As shown from middle school ELA, reducing weights on student achievement gains decreases the power to predict future student achievement gains on state tests from 0.69 to 0.63 with Model

CALCULATED WEIGHTS FOR MAXIMUM ACCURACY IN PREDICTING GAINS ON STATE TESTS

	English Language Arts			Math		
	State Tests	Observations	Student Surveys	State Tests	Observations	Student Surveys
Elementary	65%	9%	25%	85%	5%	11%
Middle	81%	2%	17%	91%	4%	5%

RELIABILITY AND ACCD ALIA

correlation with state achievement gains; it can also lower reliability and the correlation with other types of testing outcomes.

Ultimately, states, local education authorities, and other stakeholders need to decide how to weight the measures in a composite. Our data suggest that assigning 50 percent or 33 percent of the weight to state test results maintains considerable predictive power, increases reliability, and potentially avoids the unintended negative consequences from assigning too-heavy weights to a single measure. Removing too much weight from state tests, however, may not be a good idea, given the lower predictive power and reliability of Model 4 (25 percent state tests). In short, there is a range of reasonable weights for a composite of multiple measures.

Validity and Content Knowledge for Teaching

Teachers shouldn't be asked to expend effort to improve something that doesn't help them achieve better outcomes for their students. If a measure is to be included in formal evaluation, then it should be shown that teachers who perform better on that measure are generally more effective in improving student outcomes. This test for "validity" has been central to the MET project's analyses. Measures that have passed this test include high-quality classroom observations, well-designed student-perception surveys, and teachers' prior records of student achievement gains on state tests.

Over the past year, MET project researchers have investigated another type of measure, called the Content Knowledge for Teaching (CKT) tests. These are meant to assess teachers' understanding of how students acquire and understand subject-specific skills and concepts in math and ELA. Developed by the Educational Testing Service and researchers at the University of Michigan, these tests are among the newest measures of teaching included in the MET project's analyses. Mostly multiple choice, the questions ask how to best represent ideas to students, assess student understanding, and determine sources of students' confusion.

The CKT tests studied by the MET project did not pass our test for validity. MET project teachers who performed better on the CKT tests were not substantively more effective in improving student achievement on the outcomes we measured. This was true whether student achievement was measured using state tests or the supplemental assessments of higher-order thinking skills. For this reason, the MET project did not include CKT results within its composite measure of effective teaching.

These results, however, speak to the validity of the current measure still early in its development in predicting

How Can Teachers Be Assured Trustworthy Results from Classroom Observation?

Classroom observations can be powerful tools for professional growth. But for observations to be of value, they must reliably re ect what teachers do throughout the year, as opposed to the subjective impressions of a particular observer or some unusual aspect of a particular lesson. Teachers need to know they are being observed by the right people, with the right skills, and a suf cient number of times to produce trustworthy results. Given this, the challenge for school systems is to make the best use of resources to provide teachers with high-quality feedback to improve their practice.

The MET project's report Gathering Feedback for Teaching showed the importance of averaging together multiple observations from multiple observers to boost reliability. Reliability represents the extent to which results re ect consistent aspects of a teacher's practice, as opposed to other fac- tors such as observer judgment. We also stressed that observers must be well-trained and assessed for accuracy-18.84sn7(5.9(d f)-6.9(2- [(t)8.5(ho 12.5(-)-2.6(t)-7.41 f)-y 1 W.3-35.9(d t)-7.4(7.6-640.t(l3216(nt)-27.8(s t1.7(40.2(a)-6.80)-15(a)-7.6(p)-12.5(h)-mr.556 T)-6(m)

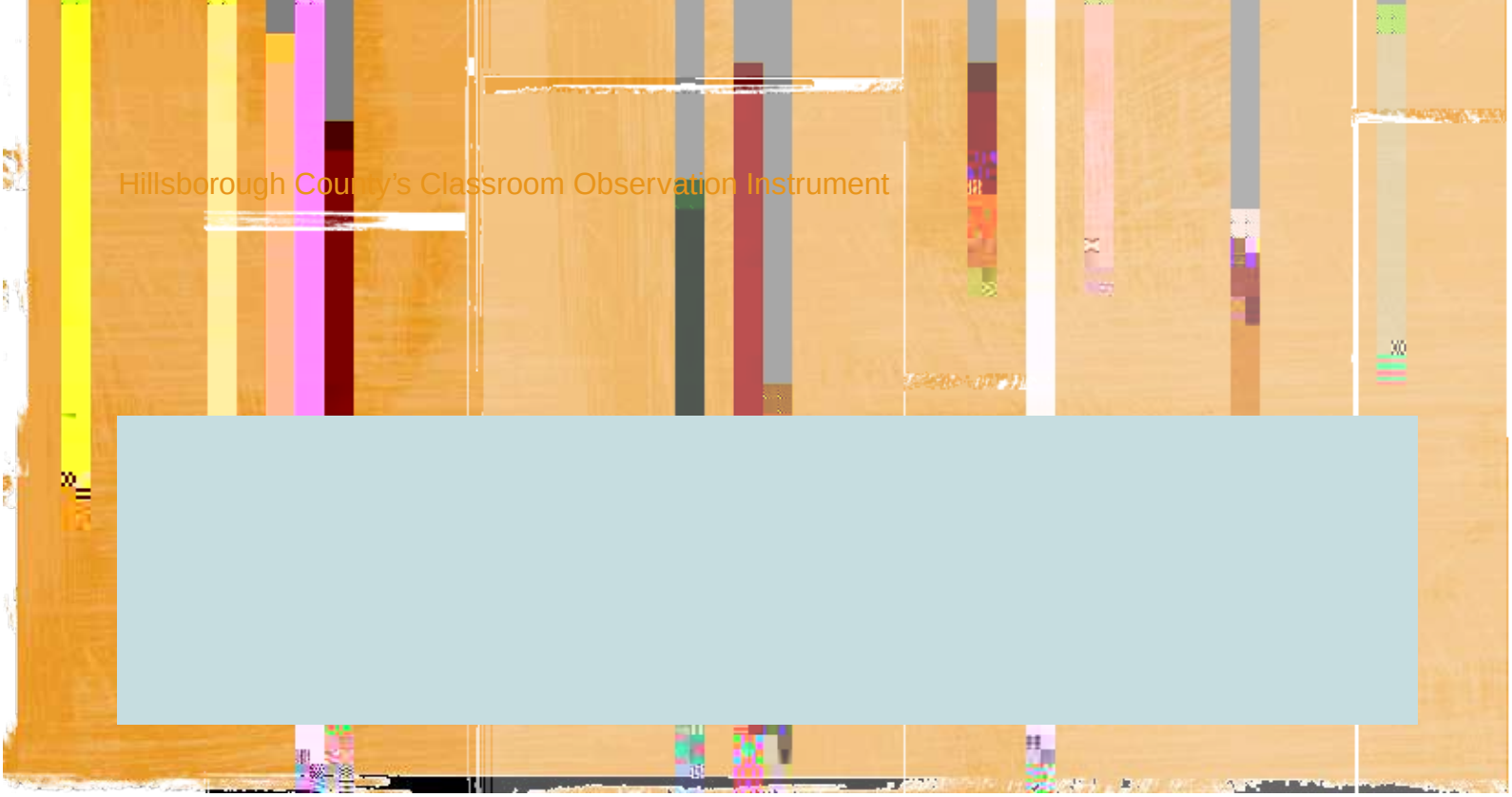

Hillsborough County's Classroom Observation Instrument

Figure 5 graphically represents many of the key findings from our analyses of those ratings. Shown are the estimated reliabilities for results from a given set of classroom observations. Reliability is expressed on a scale from 0 to 1. A higher number indicates that results are more attributable to the particular teacher as opposed to other factors such as the particular observer or lesson. When results for the same teachers vary from lesson to lesson or

things in the videos that others do, and they are not being swayed by personal biases.

If additional observations by additional observers are important, how can the time for those added observations be divided up to maximize the use of limited resources while assuring trustworthy results? This is an increasingly relevant question as more school systems make use of video in providing teachers with feedback on their practice. Assuming multiple videos for a teacher exist, an observer could use the same amount of time to watch one full lesson or two or three partial lessons.

As we move forward, MET project teachers are supporting the transition from research to practice. More than 300 teachers are helping the project build a video library of practice for use in professional development. They will record more than 50 lessons each by the end of this school year and make these lessons available to states, school districts, and other organizations committed to improving effective teaching.

This will allow countless educators to analyze instruction and see examples of great teaching in action.

Furthermore, the unprecedented data collected by the MET project over the past three years are being made available to the larger research community to carry out additional analyses, which will increase knowledge of what constitutes effective teaching and how to support it. MET project partners already are tapping those data for new studi(h 1 o)-3.2(ln 1 o)-3.23bs eruertai1.5(ini1.5(n)-158(g)-1.5(i,-1 o)-1.9(o)-143(m)]2.1ekb-3(

Bill & Melinda Gates Foundation

Guided by the belief that every life has equal value, the Bill & Melinda Gates Foundation works to help all people lead healthy, productive lives. In developing countries, it focuses on improving people's health and giving them the chance to lift themselves out of hunger and extreme poverty. In the United States, it seeks to ensure that all people—