Research Report

The Measure Matters:

Examining Achievement Gaps on Cognitively Demanding Reading and Mathematics Assessments

Marisol J. C. Kevelson

ETS POLICY EVALUATION & RESEARCH CENTER

This Policy Information Report was written by:

Marisol J. C. Kevelson
Educational Testing Service, Princeton, NJ

Policy Information Center
Mail Stop 19-R
Educational Testing Service
Rosedale Road Princeton, NJ 08541-0001
(609) 734-5212
pic@ets.org

Copies can be downloaded from: www.ets.org/research/pic

About ETS

At ETS, we advance quality and equity in education for people worldwide by creating assessments based on rigorous research. ETS serves individuals, educational institutions and government agencies by providing customized solutions for teacher certification, English language learning, and elementary, secondary and postsecondary education, and by conducting education research, analysis and policy studies. Founded as a nonprofit in 1947, ETS develops, administers and scores more than 50 million tests annually—including the *TOEFL*® and *TOEIC*® tests, the *GRE*® tests and *The Praxis Series*® assessments—in more than 180 countries, at over 9,000 locations worldwide.

Policy Information Report and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

The Measure Matters: Examining Achievement Gaps on Cognitively Demanding Reading and Mathematics Assessments

Marisol J. C. Kevelson

Educational Testing Service, Princeton, NJ

This study presents estimates of Black–White, Hispanic–White, and income achievement gaps using data from two different types of reading and mathematics assessments: constructed-response assessments that were likely more cognitively demanding and state achievement tests that were likely less cognitively demanding (i.e., composed solely or largely of multiple-choice items). Specifically, the study utilized multilevel modeling of data from over 25,000 fourth- through eighth-grade students participating in the 6-state Measures of Effective Teaching (MET) study of 2009–2010, including data from the state reading and mathematics achievement tests used in MET districts at that time and data from the Stanford Achievement Test Open-Ended Reading Assessment (SAT-9OE) and the Balanced Assessment of Mathematics (BAM). The latter two assessments, consisting entirely of constructed-response items, were selected by MET researchers to assess learning outcomes, such as those included in the Common Core State Standards, deemed more cognitively complex than those assessed by state achievement tests at the time. The investigator found that estimated Black–White, Hispanic–White, and income achievement gaps were smaller on the SAT-9OE than on state reading assessments, before accounting for other relevant factors. Estimates of Black–White and Hispanic–White mathematics achievement gaps were slightly larger using BAM data, whereas the estimated income achievement gap was slightly smaller using BAM data. In later models, prior student academic achievement and average student subject-specific prior achievement accounted for portions of these estimated achievement gaps.

Keywords Achievement gaps; racial achievement gaps; income achievement gaps; state achievement tests; constructed-response assessments; test item format; Measures of Effective Teaching study

doi:10.1002/ets2.12278

Achievement gaps, an important measure of educational equality, continue to persist in the United States between racial and ethnic minority students and White students and between students from lower and higher income households (Barton & Coley, 2009, 2010; Hemphill & Vanneman, 2011; Reardon, 2011; Vanneman, Hamilton, Baldwin Anderson, & Rahman, 2009). Defined as differences in the average standardized test scores of students from different racial groups or from families with different income levels (Lee & Burkam, 2002; Mackey et al., 2015), these gaps continue to indicate to our nation the extent of inequalities in the educational opportunities afforded to our students and, in turn, inform debates regarding the ways in which such gaps might be mitigated (García & Weiss, 2017).

Research on achievement gaps thus far has relied on data from large-scale assessments, often composed solely or largely of multiple-choice items (e.g., Fryer & Levitt, 2004; Hansen, Levesque, Quintero, & Valant, 2018; Hemphill & Vanneman, 2011; Reardon, 2011; Shannon & Bylsma, 2002; Stanford University Center for Education Policy Analysis [Stanford CEPA], n.d.; Vanneman et al., 2009). Multiple-choice assessments can and do measure a broad range of skills, including more advanced thinking skills. They also provide the opportunity to easily compare results across large groups of demographically varied students. However, multiple-choice assessments cannot measure some of the most complex skills taught by schools, such as writing and providing an answer without being reminded of it when seeing it as an option in a list (Livingston, 2009; Scully, 2017). Instead, constructed-response assessments are generally viewed as more useful measures of these and other more cognitively complex skills (Livingston, 2009; Scully, 2017; Yuan & Le, 2012).

I report here the results of a study that took advantage of the availability of large-scale data on cognitively demanding reading and mathematics constructed-response assessments as a means for providing a more fine-grained view of achievement gaps than is traditionally available from standardized assessments. Specifically, the study explores the extent

Corresponding author: M. J. C. Kevelson, E-mail: mkevelson@ets.org

to which racial and income achievement gaps measured by student scores on the constructed-response assessments administered as part of the Measures of Effective Teaching (MET) study differ from achievement gaps measured using the scores of the same students on state achievement tests. Whereas the state tests used in the six large MET school districts varied in content coverage and cognitive demand, and many consisted entirely of multiple-choice questions (Grossman, Cohen, Ronfeldt, & Brown, 2014; Polikoff, 2010; Yuan & Le, 2012), the supplemental constructed-response assessments selected by MET researchers required written responses that called upon a more advanced skill set (Ferguson & Hirsch, 2015; Grossman et al., 2014; D. McCaffrey, personal communication, April 2, 2018; White, Rowan, Alter, & Greene, 2014).

To set the stage for the current study, I first highlight the prior scholarship on achievement gaps. I then review the literature on test item format, including information on the specific tests used in this study. Next, I present the study research questions and descriptions of the methodology and results. Finally, I discuss how the study findings align with or add to the relevant literature and highlight the implications of the study findings.

Achievement Gaps

Achievement gaps have been a major national concern since the 1954 *Brown v. Board of Education* Supreme Court decision declared that segregated schools are unequal and unconstitutional because disparities in school quality lead to different school outcomes (Legal Information Institute, n.d.). In 1966, the Coleman Report documented the continuing racial inequity in student achievement, becoming the first in a long line of empirical studies on achievement gaps (Coleman et al., 1966; Jencks & Phillips, 1998; Jones, 1984; National Center for Education Statistics [NCES], 2018).

Research also documents income achievement gaps, another measure of inequity in student achievement, going back decades (Reardon, 2011). Concerns over racial and socioeconomic disparities in school quality persist, despite a plethora of school reform policies designed to improve the equity of educational opportunities across the United States (Barton & Coley, 2009, 2010; Darling-Hammond, 2014; Duncan & Magnuson, 2011; Farkas, 2011; Fryer & Levitt, 2004).

Income achievement gaps are linked to racial achievement gaps because race has a long history of being associated with differences in employment opportunities, earnings, and educational opportunities (Grodsky & Pager, 2001; Jencks & Phillips, 1998; Pager & Shepherd, 2008; Reardon, 2011). As a result, racial minorities more often earn lower incomes and are overrepresented among individuals living below the federal poverty line (Kaiser Family Foundation, 2017). It is not surprising that racial achievement gaps are highly correlated with state levels of racial socioeconomic disparities (Reardon, 2011). The parents of Black and Hispanic children tend to have lower incomes and lower levels of educational attainment than parents of White children, and family socioeconomic resources are strongly related to educational outcomes because more affluent and more educated families tend to provide more educational opportunities for their children (Lareau, 2011; Organisation for Economic Co-operation and Development, 2016; Phillips, 2011; Stanford CEPA, n.d.). Racial and income achievement gaps may also be influenced by socioeconomic differences in social expectations, mobility rates, health concerns, chronic stress, language exposure, and financial assets (Ackerman & Brown, 2010; Duncan & Brooks-Gunn, 1997; Lareau, 2011; Rothstein, 2004). School segregation by socioeconomic strata continues to be an issue, given that students tend to learn less in classes composed largely of children from low-income families (Duncan & Magnuson, 2011; Farkas, 2011; Raudenbush, Marshall, & Art, 2011) and that teacher commitment, parental involvement, and student achievement all tend to be low in schools in low-income and high-crime neighborhoods (Kirk & Sampson, 2011). However, racial achievement gaps persist even in states where the racial socioeconomic disparities are almost nonexistent—for example, states with small Black or Hispanic populations (Reardon, 2011). Moreover, school quality remains a salient predictor of achievement gaps even after socioeconomic differences are accounted for (Fryer & Levitt, 2004).

Racial Achievement Gaps

Although the good news is that racial achievement gaps have decreased over time, progress has been uneven and has stalled in recent years, even reversing for a while in the 1990s (Barton & Coley, 2010; Lee & Burkam, 2002; Stanford CEPA, n.d.). The bad news is that racial achievement gaps are persistent and still quite large. Although data from the National Assessment of Educational Progress (NAEP) demonstrated a 30–40% reduction in the national Black–White and Hispanic–White reading and mathematics achievement gaps from the 1970s to 2012, these gaps remain sizable,

ranging from 0.5 to 0.9 *SD*, for Grades 4, 8, and 12 (Hemphill & Vanneman, 2011; Stanford Center for Education Policy Analysis, n.d.; Vanneman et al., 2009).

State-level NAEP data show that racial and ethnic reading and mathematics achievement gaps within states have an even larger range of 0.3 *SD* to 1.5 *SD*, indicating that some states are doing better than others at closing or minimizing racial achievement gaps (Hansen et al., 2018; Stanford CEPA, n.d.). In fact, racial achievement gaps have narrowed in some states while they have widened in others (Stanford CEPA, n.d.). Within the six states participating in the MET study, Black – White achievement gaps ranged from 0.4 *SD* to 1.0 *SD* in reading and 0.6 *SD* to 1.0 *SD* in mathematics on the main NAEP test administered to states in 2011. Hispanic – White achievement gaps ranged from 0.2 *SD* to 1.0 *SD* in reading and mathematics on the same tests (Stanford CEPA, n.d.).

States also differ in the relative direction of racial achievement gaps. In some states, achievement gaps are large because White students score particularly high, not because Black or Hispanic students score especially low. In other states, the opposite is true, and Black or Hispanic students score especially low, so they perform below even a relatively low-achieving White student population. Futhermore, despite the correlation between state racial and socioeconomic disparities and racial achievement gaps, some states with similar levels of socioeconomic disparities have substantially different achievement gaps. This suggests that other factors — including the availability and quality of early childhood education, the quality of public schools, patterns of residential and school segregation, and state educational and social policies — may play important roles in reducing or exacerbating racial achievement gaps (Duncan & Magnuson, 2011; Farkas, 2011; Fryer & Levitt, 2004; Stanford CEPA, n.d.). The tendency within the United States to favor local-level school control or policies and practices has contributed to variations in the size of achievement gaps within and across states.

Given that states with high per-pupil spending or Common Core State Standards implementation are not doing noticeably better than their counterparts (Hansen et al., 2018), other state-level educational policies may be more influential. This is true not only regarding racial achievement gaps, but also regarding state-level differences in income achievement gaps. Overall, NAEP data show that across states Black – White achievement gaps tend to be largest, income achievement gaps measured using free and reduced-price lunch eligibility tend to be smaller, and Hispanic – White achievement gaps tend be smaller still (Hansen et al., 2018). In addition, it appears that states with relatively small income achievement gaps tend to have relatively small Black – White and Hispanic – White gaps.

Income Achievement Gaps

It seems that the achievement gap between low- and high-income families has widened as the income gap between these families has grown (Reardon, 2011). In fact, the income achievement gap is approximately 30 – 40% larger among children born in 2001 than among children born 25 years earlier, and it may have been increasing for at least 50 years. Moreover, achievement gaps between those at the 90th percentile and those at the 10th percentile of the income distribution (the "90/10 gap") are even larger than racial and ethnic achievement gaps, at 1.25 *SD* in reading and approximately 1.30 *SD* in mathematics in 2008 (Reardon, 2011).

Another way to track income achievement gaps is to compare the achievement test scores of students eligible and ineligible for free or reduced-price lunches through the National School Lunch Program,[1] a commonly used indicator of being from a low-income family (Domina et al., 2017). Research using this approach indicates that the national free and reduced-price lunch achievement gap decreased by only 0.02 *SD* between 2003 and 2017, according to NAEP data (Hansen et al., 2018). However, the 90/10 gaps studied by Reardon (2011)[2] may be a better measure of income achievement gaps than free and reduced-price lunch eligibility. For example, students living in households with incomes just below the eligibility threshold may have many more advantages than those living in households at the bottom of the income distribution who are also eligible for free lunch. Similarly, students from households with incomes just above the income eligibility threshold may have far fewer advantages than wealthy students who are also ineligible for free lunch (Hansen et al., 2018). However, some research has shown that free and reduced-price lunch eligibility captures disadvantage on test scores above and beyond household income (Domina et al., 2017). Moreover, whereas the 90/10 gaps focus only on the extremes of the income distribution, the free and reduced-price lunch gap is a broader measure inclusive of Americans at all income levels.

Thus, whereas there is clear evidence of both the decline and the persistence of racial achievement gaps, there is mixed evidence on the growth or decline of income achievement gaps, depending on the measure used. Yet both racial and

income achievement gaps remain important indicators of inequalities in society and schooling. Further research is needed regarding the measurement of gaps and solutions to reduce the racial and income inequalities that they indicate.

Test Item Format and Achievement Gaps

One potential avenue of study is the extent to which achievement gaps vary when different types of test items are used. Gender achievement gaps have been shown to be strongly associated with test item format, such that 25% of state-level variance in gaps is explained by the proportion of multiple-choice versus constructed-response test items (Reardon, Kalogrides, Fahle, Podolsky, & Zárate, 2018). Although there is a large body of existing evidence on racial and ethnic test score differences by test item format, taken together the results are inconclusive. And there is only limited research on the extent to which income achievement gaps may vary by test item format.

Item Format and Racial Achievement Gaps

Some studies have found that racial minority students score lower than White students on constructed-response items, multiple-choice items, or both. In a study using achievement test data from Washington state, Shannon and Bylsma (2002) compared item-level scores across racial subgroups and found that achievement gaps between Black, Hispanic, and Native American students and White students were larger on Grades 4, 7, and 10 constructed-response reading and mathematics test items than multiple-choice items. The only exception was Grade 4 reading, for which the gaps between White students and the other three groups were roughly the same for both types of items (Shannon & Bylsma, 2002). Prior studies have found that racial achievement gaps are as large or larger on constructed-response items than on multiple-choice items (Bond, 1995; Dimitrov, 1999; Feinberg, 1990; Linn, Baker, & Dunbar, 1991).

Other studies have found that racial achievement gaps are smaller on constructed-response items, although across all studies minorities tend to fare more poorly than White students on items of both types. In a study comparing the scores of Black and White individuals on constructed-response and multiple-choice assessments of the same construct, Arthur, Edwards, and Barrett (2002) found a 39% reduction in subgroup differences on the constructed-response version of the assessment compared with the multiple-choice test. Earlier studies also found smaller racial achievement gaps on constructed-response items relative to multiple-choice items, although gaps existed on items of both types (Badger, 1995; Lawrence, Lyu, & Feigenbaum, 1995; O'Neil & Brown, 1998). Another study, using data from 1994 to 1997 Michigan state tests for a small sample of schools, found no patterns of ethnic differences by test item format (DeMars, 2000).

Only one study has found that minority students outperform White students on constructed-response items. In a study using Washington state reading achievement test data for Grades 4, 7, and 11 from 1997 to 2001, Taylor and Lee (2011) found that, when subgroup performance differed on items, White students outperformed minority students on multiple-choice items, and Black, Hispanic, and Asian students outperformed White students on constructed-response test items. After exploring variations by item content and complexity and accounting for scoring processes, Taylor and Lee concluded that most items flagged for racial and ethnic differences were items that measured higher order reading skills and that the constructed-response items were scored in such a way that unique explanations were given credit, whereas this was not the case for multiple-choice items.

Item Format and Income Achievement Gaps

In contrast to the relatively large amount of research on racial differences in test performance by test item format, there appears to be only one prior study on variations in income achievement gaps by item format (Wright et al., 2016). Using data on a localized sample of college students, the authors of this study found that students from households with low socioeconomic status fared less well on exams as the percentage of constructed-response items increased.

Item Format and Scoring

There are well-known challenges inherent in scoring large numbers of open-ended responses, including rater error (Livingston, 2009; Patz, Wilson, & Hoskens, 1997; Scully, 2017), and there are, relatedly, many advantages of multiple-choice assessments (Livingston, 2009; Scully, 2017). As a result, large-scale assessments used to measure achievement gaps, such

as NAEP and state achievement tests, tend to include either a mix of open-ended and multiple-choice test items or only multiple-choice test items (NAEP, 2019; Yuan & Le, 2012). This makes sense, given the many advantages of multiple-choice tests when large numbers of students must be tested. These advantages include the fact that a sizable number of items can be completed in a short amount of time, enabling the testing of a broad range of concepts and providing a reasonably good sample of the examinees' knowledge (Livingston, 2009). Machine scoring of items, which is faster and less expensive than human scoring, is also more straightforward for multiple-choice items than for constructed-response items (Livingston, 2009). Some even maintain that multiple-choice items can be substituted for constructed-response items based on studies showing a high level of agreement between scores on multiple-choice and constructed-response tests (Lukhele, Thissen, & Wainer, 1994). However, others note that these studies' use of the same students may mask gender differences and the extent to which scores change over time (Livingston, 2009).

Item Format and Cognitive Demand

Despite claims to the contrary, multiple-choice tests can be cognitively demanding when more of the most rigorous items — those requiring comprehension, application, and analysis skills — are included (Scully, 2017). However, as one psychometrician noted, "Many skills schools teach are too complex to be measured effectively with multiple-choice questions" (Livingston, 2009, p. 1). Examples include writing essays, communicating and organizing thoughts clearly in writing, and providing an answer without being reminded of it when seeing it as an option in a list. Skills that might be called upon by a constructed-response assessment include remembering, without reminders, the procedure for a science experiment; explaining the logical error in a persuasive text; generating new ideas; and critiquing existing ideas (Livingston, 2009; Scully, 2017).

Constructed-response assessments thus have several advantages over multiple-choice tests. They have the potential to measure the highest level thinking skills, including divergent thinking, evaluation, and creation skills. They can offer better measures of writing skills and opportunities for extended thinking because they can include complex questions and open-ended problems requiring interpretations of results, formulations of conclusions, and explanations of the processes used to solve the problem (Darling-Hammond, 2017; Darling-Hammond & Adamson, 2010; Madaus & O'Dwyer, 1999; Stecher, 2010). The higher level thinking skills that constructed-response assessments have the potential to measure are viewed by some as critical for the nonroutine jobs increasingly available in the "knowledge economy" and, thus, are viewed as important for college and career success and, arguably, for participation in a democratic society (Darling-Hammond, 2017; Darling-Hammond & Adamson, 2010; Liu, Frankel, & Roohr, 2014).

Beyond having the potential to document higher order thinking skills, constructed-response assessments can be used formatively to collect information on students' weaknesses and strengths and enable teachers to tailor instruction appropriately (Black & Wiliam, 1998; Darling-Hammond & Adamson, 2010). Whereas data from multiple-choice test items can certainly be used formatively, once again, constructed-response assessments have the advantage in their potential to assess a broader range of skills (Scully, 2017). Moreover, formative assessments that require higher order thinking skills may foster conceptual understanding and even make students more likely to use higher versus lower level skills in their schoolwork (Jensen, McDaniel, Woodard, & Kummer, 2014; Scully, 2017). The MET supplemental assessments, in particular the Balanced Assessment of Mathematics (BAM), were designed to provide these types of formative and learning experiences for students and teachers.

MET Constructed-Response Assessments

Whereas constructed-response assessments may be used in local, state, and even national assessments, the challenges inherent in scoring large numbers of constructed-response assessments mean that large-scale data on solely constructed-response assessments are not common. The MET study, which took place in 2009–2010 and 2010–2011 (and is discussed in more detail below), provided a large number of valid and reliable constructed-response test scores for six large metropolitan areas in the United States, along with scores on state achievement tests used in MET districts. These data therefore provide an opportunity to contrast scores from constructed-response assessments with those from more traditional multiple-choice tests or tests with mixed-item formats (White et al., 2014).

In addition to collecting state achievement test score data for students taught by participating teachers, MET researchers opted to administer constructed-response reading and mathematics assessments to each of these students.

For this purpose, they selected the Stanford 9 Open-Ended Reading Assessment (SAT-9OE), which asks students to write short-answer responses to questions testing their comprehension (Pearson Education, 2018), and the BAM, which was designed to measure students' conceptual understanding of math topics (Balanced Assessment in Mathematics Project, 1995). Both assessments were selected because MET researchers ascertained that their items are cognitively demanding; their content is reasonably well aligned to classroom curricula; and there is strong evidence of their validity, reliability, and fairness for members of different groups of students (Kane & Cantrell, 2010). MET researchers had an interest in testing whether findings for accountability exams, often used to evaluate teacher performance, also apply to exams not used for accountability. They also sought to understand whether the value-added scores used to evaluate teachers might differ when calculated using more cognitively demanding tests (Ferguson & Hirsch, 2015). Because these supplemental assessments were "low stakes," it was expected that scores on the MET supplemental assessments should not be influenced by "teaching to the test," whereas this was viewed as a distinct possibility for high-stakes state achievement tests (Kane & Staiger, 2012; Shepard, 1997).

The MET study findings indicated that teachers who were promoting gains on state assessments may have also been promoting deeper conceptual understanding among their students based on students' scores on the BAM and SAT-9OE (Kane & Cantrell, 2010). However, the moderate correlations between state math test scores and BAM scores (.45) and between state reading test scores and SAT-9OE scores (.46) indicated that state test scores may be driven in part by aspects of teacher performance that are specific to the state test and that may not generalize to other student outcomes of interest (Kane & Staiger, 2012), such as higher order thinking skills.

Highlighting the greater potential of the SAT-9OE to assess a broad range of reading skills as compared to state reading tests used in MET districts, MET researchers also found teacher effects to be similar on the SAT-9OE, the BAM, and state mathematics tests, but not on state reading tests (Kane & Staiger, 2012). Moreover, other researchers comparing the SAT-9OE, state reading tests, and the Protocol for Language Arts Teaching Observation (PLATO) found that the SAT-9OE is more instructionally sensitive to the PLATO factor of Cognitive and Disciplinary Demand than are the state tests used in the MET study (Grossman et al., 2014).

The Current Study

Both race and income achievement gaps are evident as early as preschool and may have lifetime consequences stemming from wide disparities in high school and college completion, which perpetuate differences in occupations, employment, and earnings (Duncan & Magnuson, 2011; Jencks & Phillips, 1998; Murnane & Levy, 1996; Sharkey, 2013). Because different types of test items can measure different types of skills, achievement gaps measured using different test item formats may provide a more comprehensive understanding of gaps on a range of skills. In this study, I set out to document the extent to which achievement gaps might differ when measured using data from reading and mathematics constructed-response assessments compared to data from state reading and mathematics assessments. I use multilevel modeling of MET test score data and student demographic data for Grades 4–8 to address the following research questions:

1. To what extent do scores on the MET constructed-response reading and mathematics assessments vary by student race or ethnicity?
2. To what extent do these scores vary by family income (e.g., eligibility or ineligibility for free or reduced-price lunch)?
3. How do racial or income achievement gaps in scores on the MET constructed-response reading and mathematics assessments compare to achievement gaps on state reading and mathematics assessments used in MET districts?

Method

Data and Sample

The MET study collected data on over 150,000 students taught by 2,756 teachers in 317 schools in six large US school districts in 2009–2010 and 2010–2011. Opportunity sampling was used to recruit districts; then elementary, middle, and high schools within each district; then teachers of reading, mathematics, and other subjects in Grades 4 through 9. The samples selected from the larger data set for the current study were created separately for reading and mathematics. Because teachers participating in the MET study were largely asked to submit data for only one subject area, and relatively few students had test score data for both reading and mathematics tests, samples had to be created separately for

each subject in order to maximize the availability of data for analyses (White et al., 2014). The reading sample included students in Grades 4–8 linked to MET teachers who exclusively taught reading or taught both reading and mathematics (i.e., elementary school teachers); likewise, the mathematics sample included students in Grades 4–8 linked to teachers who exclusively taught mathematics or taught both mathematics and reading.[3] The reading sample included 27,143 students in Grades 4 through 8 taught by 879 teachers, and the mathematics sample included 25,233 students in the same grades, taught by 817 teachers.[4] The study samples included only data collected in 2009–2010 because of teacher attrition in 2010–2011. Further, because MET tracked teachers, not students, over time, it was not possible to use student data collected by MET researchers[5] from both years to track changes over time.

Measures

SAT-9OE and BAM Constructed-Response Assessments

In addition to collecting district administrative data on teacher and student demographics and state achievement test results, MET researchers used the SAT-9OE and the BAM to collect supplemental data on students in MET classrooms. Prior to selecting the SAT-9OE and BAM, MET researchers conducted an alignment study to ascertain the extent to which the assessment content aligned with the state standards for each of the six participating districts and confirmed that the content of both assessments was adequately aligned with state standards (D. McCaffrey, personal communication, April 2, 2018).

The SAT-9OE contains nine open-ended questions and takes 50 min to complete.[6] The primary difference between the SAT-9OE and the state reading achievement tests used in MET states at the time of the study is the former's exclusive use of open-ended items tied to lengthy reading passages. Each form of the assessment consists of a narrative reading selection followed by nine comprehension questions. Students are required not only to answer the questions, but also to explain their answers (Kane & Cantrell, 2010). Rather than using the standard form of the SAT-9, which also includes multiple-choice items, MET researchers modified the assessment by including only the Open-Ended Reading Assessment. This decision was made because the Open-Ended Reading Assessment focuses exclusively on higher order thinking skills, and as noted above, the MET researchers were interested in using supplemental assessments that are more cognitively demanding than achievement tests used in MET states. During Year 1 of the MET study, 2009–2010, the focal year of this analysis, 75% of students in the fourth- through eighth-grade reading sections consented to take, and completed, the SAT-9OE (White et al., 2014). The tests were each rated by a single rater using Pearson's standard commercial practices.[7] Scaled scores ranged from 425 to 838, with a mean of 624 and standard deviation of 42.

The BAM assessment used by MET researchers includes five tasks and requires 50–60 min to complete. Tasks require students to solve real-world mathematics problems and to explain how they arrived at their solutions. Because of the small number of tasks on each test form, MET researchers were concerned about the content coverage in each teacher's classroom. As a result, they used three different forms of the BAM in each classroom. In comparison to many other assessments, BAM is considered to be more cognitively demanding and measures higher order reasoning skills by using question formats that are quite different from those in most state mathematics achievement tests used in MET districts during the years of the MET study (Kane & Staiger, 2012). The BAM was developed as part of the Mathematics Assessment Project (MAP) undertaken by an international collaborative of curriculum developers known as the Mathematics Assessment Resource Service. The BAM is now known as the Classroom Challenges, a set of "formative assessment lesssons" (MAP, 2015; D. Foster, personal communication, February 28, 2019). Thus, the BAM assessment used in the MET study was a version based on decades of refinement and field testing on a normed sample, after the initial development of the BAM in the 1990s. The original BAM developers noted that their tasks differ from traditional standardized tests in their ability to assess problem-solving abilities and provide information on how a student reasons, communicates mathematically, and makes connections across mathematical content (Balanced Assessment in Mathematics Project, 1995). BAM tasks require students to create a plan, make a decision or solve a problem, and then justify their thinking. More recent evaluations of the BAM/Classroom Challenges tasks note their emphasis on higher order thinking skills including analysis, synthesis, justification, and reflection and their alignment with and support of teaching to nurture higher order thinking skills and address the Common Core State Standards (Inverness Research, 2014; Research for Action, 2015). Scores on the current version of the BAM assessment are also associated with scores on the California state Smarter Balanced Assessment System (D. Foster, personal communication, February 28, 2019). BAM may even be

more instructionally sensitive to the effects of this type of reform-oriented instruction than a more traditional test (Kane & Staiger, 2012).

BAM tasks are scored on a 7- to 12-point scale, depending on the task, using rubrics developed for each task; scores are summed across the five administered tasks. Most of the BAM assessments administered for the MET study were scored by one rater and then checked by a second rater; a subsample of assessments was double-scored to check for accuracy and consistency (D. Foster, personal communication, February 28, 2019). Raw scores ranged from 0 to 40 points for the MET administration of the BAM, with a mean of 16.43 points and an *SD* of 9.38 points. During Year 1 of the MET study, 79% of students in the fourth- through eighth-grade math classes completed the BAM (White et al., 2014).

Although interrater reliabilities were not available for the scoring of the BAM or the SAT-9OE for the MET study specifically, they are available for BAM for a prior sample. Interrater reliabilities for an audited subsample of ratings of BAM scores obtained from a normed sample over 3 years prior to the MET study (2003 through 2005) ranged from 74% to 92% rater agreement within 1 point, with score correlations of approximately .99 (Mathematics Assessment Collaborative, 2003, 2004, 2005).Whereas the SAT-9 reliability and validity study includes measures of reliability focused on internal consistency — with Cronbach's alpha reliability coefficients ranging from .78 to .88 for Grades 4 through 8 — it does not include measures of interrater reliabilities for the SAT-9OE (Harcourt Brace Educational Measurement, 1996).

Although the Webb's Depth of Knowledge (DOK) tool was not explicitly used to rate the cognitive demand of the BAM and SAT-9OE items, it is reasonable to assume that the items may require cognitive skills extending into DOK Level 3 and even Level 4. The DOK scale reflects the full range of cognitive skills, from lower level thinking skills at Levels 1 and 2 to higher order thinking skills at Levels 3 and 4 (Synergis Education, 2018). Level 1 includes the skills of recall and recognition; Level 2 includes basic reasoning and comparisons; Level 3 involves more complex reasoning such as analysis, planning, and justification; and Level 4 is development of thinking and ideas, often over an extended period.

As noted above, BAM tasks require students to create a plan, make a decision or solve a problem, and then justify their thinking. The tasks assess problem-solving abilities and require analysis and synthesis skills characteristic of DOK Level 3, in addition to the extended thinking characteristic of test items at DOK Level 4 (Balanced Assessment in Mathematics Project, 1995; D. Foster, personal communication, February 28, 2019). Also as noted above, the SAT-9OE consists of only open-ended items tied to lengthy reading passages. The items assess comprehension and require higher order thinking in the form of analysis of text and justification for selected answers (Kane & Cantrell, 2010). The SAT-9OE technical documentation noted that all open-ended questions measure "thinking skills," which include:

> The ability to analyze and synthesize information; to classify and sequence information; to compare and contrast information; to evaluate information in order to determine cause and effect, fact and opinion, relevant and irrelevant; and to interpolate and/or extrapolate beyond information in order to draw conclusions, make predictions, and hypothesize. (Harcourt Brace Educational Measurement, 1996, p. 23)

It seems that a focus on measurement of these skills could indicate a DOK level as high as 3.

Both BAM and SAT-9OE scores were standardized within each grade to account for variations in test forms and subsequently for the overall sample. The final variables were standardized across all students to enable interpretation of coefficients in effect sizes.

2009–2010 State Test Scores

District administrative data files provided to MET researchers included student scores on state achievement tests in reading and mathematics for the 2 years of the MET study and the 2 prior years. I used state reading and mathematics achievement test scores from 2009 to 2010, the focal year of the study, as outcome measures in multilevel models predicting achievement gaps on state reading and mathematics tests. State reading and mathematics test scores from 2008 to 2009 were used as a control for prior achievement in models predicting achievement gaps on state reading and mathematics tests and in models predicting achievement gaps on the MET constructed-response assessments.

Documentation from state departments of education and the NCES (2011a) revealed that the 2008–2009 or 2009–2010[8] reading and mathematics tests administered in Grades 4 through 8 in the six MET states were primarily multiple choice (see Appendix A). In terms of cognitive complexity, one study noted that most items on three of the six 2009–2010 MET state tests, Colorado, New York, and Texas, were rated at a DOK level of 1 or 2 on a 4-point scale (Yuan

Table 1 Percentages of Students Scoring at or Above *Proficient* on Achievement Tests in Measures of Effective Teaching States Versus National Assessment of Educational Progress, Grade 4 and 8 Reading and Mathematics

	Reading				Mathematics			
	Grade 4		Grade 8		Grade 4		Grade 8	
State	At or above state proficiency standards	At or above NAEP proficiency level	At or above state proficiency standards	At or above NAEP proficiency level	At or above state proficiency standards	At or above NAEP proficiency level	At or above state proficiency standards	At or above NAEP proficiency level
Colorado	87%	40%	88%	32%	91%	45%	81%	40%
Florida	74%	36%	54%	32%	75%	40%	66%	29%
North Carolina	69%	32%	66%	29%	81%	43%	80%	36%
New York	77%	36%	68%	33%	87%	40%	80%	34%
Tennessee	90%	28%	93%	28%	90%	28%	90%	25%
Texas	84%	28%	94%	27%	85%	38%	83%	36%

Note. NAEP = National Assessment of Educational Progress. Data are from 2009 State Mapping Analysis by NAEP (2011). Retrieved from https://nces.ed.gov/nationsreportcard/studies/statemapping/2009_naep_state_table.aspx

& Le, 2012). Technical documentation from another MET state, Florida, showed that the percentages of highly complex reading assessment items were 10–20% for Grade 4, 15–25% for Grades 5–7, and 20–30% for Grade 8. Similarly, the percentages of highly complex mathematics assessment items were 5–15% for Grade 4, 20–30% for Grade 5, 10–20% for Grades 6–7, and 20–30% for Grade 8 (Florida Department of Education, 2009). Technical documentation for the Tennessee and North Carolina 2009–2010 tests either did not indicate the proportions of items developed to address each level of cognitive complexity or was not available.

Further evidence of the lower cognitive demand of state achievement tests can be found in a comparison of proficiency rates on the 2009 achievement tests for the six MET states and the main NAEP of 2009. As shown in Table 1, across all six states and in both Grades 4 and 8, proficiency rates were higher on state reading and mathematics achievement tests than on the NAEP test of the same subject. Given the well-documented rigor of the NAEP (National Assessment Governing Board, 2017a, 2017b), this trend provides further indication that achievement tests in the six MET states may have been less cognitively demanding than the BAM and SAT-9OE administered by MET researchers in 2009–2010.

In this study, both reading and mathematics scale scores were standardized, first within grade and then within district, and subsequently for the overall sample, to adjust for differences in assessment scales between different grades and districts. Although this approach does not address differences between state assessments in content and complexity noted by prior researchers (Barton & Coley, 2009), fortunately, as noted above, prior to selecting the BAM and SAT-9OE MET, researchers conducted an alignment study to verify that the content of these tests aligned with the state standards for each of the six participating districts (D. McCaffrey, personal communication, April 2, 2018). Thus, differences by content are assumed to be minimal, although some differences may still exist given variations in state learning standards at the time of the MET study.[9] The final variables were standardized across all students to enable interpretation of coefficients in effect sizes.

Race and Income

Dependent variables included measures of student race/ethnicity and eligibility for free or reduced-price lunch.[10] Dichotomous 0/1 indicators were used for each of the race/ethnicity categories included in the study (Black, Hispanic, Asian, other race, and White). The Black, Hispanic, Asian, and other race indicators were entered into models for analysis, whereas the White variable was omitted as the comparison category. As per the regulations of the National School Lunch Program (see Footnote 1), students are eligible for free lunch if their family income is below 130% of the federal poverty line or for reduced-price lunch if their family income is below 185% of the federal poverty line (Domina et al., 2017; U.S. Department of Agriculture Food and Nutrition Service, 2018). During the 2009–2010 school year, when the federal poverty line was $22,050 per year for a family of four, families would have had to earn less than $40,793 per year to qualify for reduced-price lunch or less than $28,665 per year to qualify for free lunch (U.S. Department of Agriculture Food and

Table 2 Student and Teacher Characteristics: Descriptive Statistics

Characteristics	ELA sample (N = 27,143) Means/proportions	*SD*	Math sample (N = 25,233) Means/proportions	*SD*
Student characteristics				
White	25%	0.43	25%	0.43
Black	37%	0.48	37%	0.48
Hispanic	29%	0.46	29%	0.45
Asian	7%	0.26	6%	0.24
Other race	3%	0.16	3%	0.17
Free and reduced price lunch (FRL) student	53%	0.50	50%	0.50
FRL White	8%	0.28	6%	0.24
FRL Black	18%	0.38	17%	0.38
FRL Hispanic	23%	0.42	22%	0.42
FRL Asian	4%	0.19	3%	0.17
FRL other race	1%	0.12	1%	0.11
Male	50%	0.50	50%	0.50
Gifted student	12%	0.32	9%	0.28
Special education student	8%	0.27	8%	0.28
English language learner	13%	0.34	13%	0.34
Grade	2.90	1.37	0.13	0.34
Classroom characteristics				
Years teacher experience teaching in district	7.02	5.74	5.06	6.40
Teacher holds master's or higher degree	41%	0.49	41%	0.45

Note. ELA = English Language Arts.

Nutrition Service, 2009). For this study, the free and reduced-price lunch variable was available in the MET database as a dichotomous indicator of student household eligibility for free or reduced-price lunch.

Control Variables

Measures of student and teacher demographic characteristics were used to hold constant differences in student background and teacher qualifications that might also influence student achievement on state tests and performance on the BAM and SAT-9OE. Student-level control variables included student grade, gender, eligibility for gifted or special education services, and English language learner status. With the exception of grade, all student variables were available in the MET database as dichotomous indicator variables. Teacher characteristics included years of experience teaching in the district[11] and possession of a master's or higher degree, because some evidence indicates that these teacher characteristics influence teacher effectiveness (Bryk, Sebring, Allensworth, Easton, & Luppescu, 2010). Possession of a master's or higher degree was included as a dichotomous indicator variable based on the MET variable on teacher degree.

To account for the contextual effect of exposure to peers of differing ability levels, I created a measure of average classroom ability by aggregating state reading or mathematics test scores, respectively, to the classroom level. This measure used the state test scores that were standardized at the grade level, district level, and overall. The classroom-level version was standardized once again to enable interpretation in units of effect size.

Sample Characteristics

As Table 2 shows, the study samples included sizable proportions of Black and Hispanic students; this is not surprising, given the MET study's focus on large urban metropolitan areas. Half the students in the mathematics sample and just over half the students in the reading sample qualified for free or reduced-price lunch. Black and Hispanic students more often qualified for free or reduced-price lunch than White and Asian students. In the sample, 9–12% of students were considered "gifted," 8% qualified for special educational services, and 13% were English language learners.

Approximately 41% of teachers in both the reading and mathematics samples held at least a master's degree. Teachers in the reading sample had over 7 years of experience teaching in their districts, on average, whereas teachers in the mathematics sample had just over 5 years of experience teaching in their districts, on average.

Multilevel Models

As I previously noted, I used SAT-9OE and BAM scores as outcome measures in multilevel linear regression models. State reading and mathematics achievement tests were also used as outcome measures in a second set of models predicting achievement gaps on state tests for the purpose of comparison with gaps in scores on the SAT-9OE and BAM. Multilevel modeling is ideal for studies of educational phenomena because it enables researchers to address nested data by simultaneously accounting for variations at each level (Raudenbush & Bryk, 2002). I used the Hierarchical Linear Modeling software program, version 7.01 (Raudenbush, Bryk, & Congdon, 2013), to specify multilevel models with students nested within teachers. After specifying unconditional models to partition the variance in the four outcome measures, using a one-way random effects analysis of variance, I specified a series of two-level intercepts-as-outcomes models in which student- and teacher-level control variables were added in steps to base models, including free or reduced-price lunch status and student race indicator variables.

In all models, continuous variables were standardized, and categorical variables were coded as dichotomous indicators, as described previously, so that slope coefficients were in units of effect size, using the metric of *SD* units. Dichotomous indicator variables, including variables representing student race/ethnicity, free or reduced-price lunch eligibility, gender, eligibility for services for gifted or English language learner students or special educational services, and teacher possession of an advanced degree were entered into models uncentered. Continuous measures of prior achievement and average prior achievement were entered uncentered because they were standardized to a mean of 0 and *SD* of 1. Student grade level and teachers' years of experience teaching in the district were grand mean centered such that they represented grade level and average years of experience teaching in the district, respectively (Enders & Tofighi, 2007).

The final within-teacher model for the SAT-9OE and BAM scores, respectively (Y_{ij}), of the *i*th student of the *j*th teacher was

$$Y_{ij} = \beta_{0j} + \beta_{1j}\left(\text{Race/Ethnicity}_{ij}\right) + \beta_{2j}\left(\text{FRL}_{ij}\right) + \beta_{3j}\left(\text{Gender}_{ij}\right) + \beta_{4j}\left(\text{State test score}_{ij}\right) + \beta_{5j}\left(\text{Grade}_{ij}\right) + \beta_{6j}\left(\text{Gifted}_{ij}\right) + \beta_{7j}\left(\text{Special Education}_{ij}\right) + \beta_{8j}\left(\text{ELL}_{ij}\right) + e_{ij},$$

where β_{0j} is the intercept, β_{1-8j} are slopes, and e_{ij} is the student-specific random error, and where FRL is free or reduced-price lunch and ELL is English language learner. The student-level equation intercept was allowed to vary freely from student to student to model within-teacher variation. All other student-level slopes were constrained to represent average student-level variable estimates. The final within-teacher model for models predicting state reading and mathematics achievement test scores, respectively, were the same.

The final between-teacher model, for all four test scores, representing the within-teacher intercept, is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\left(\text{Yrs of Experience Teaching in District}_{j}\right) + \gamma_{02}\left(\text{Master's or Higher}_{j}\right) + \gamma_{03}\left(\text{Avg Class Reading or Math Score}_{j}\right) + u_{0j},$$

where γ_{00} is the intercept, γ_{01-03} are slopes, and u_{0j} is the teacher-level random error.

Missing data on outcome measures and covariates were addressed using multiple imputation via a chained equations routine.[12] Multiple imputation reduces bias caused by missing data through a missing values prediction process that preserves important parts of, and connections within, the data distribution (Allison, 2001; Schafer & Olsen, 1998). In data sets with imputed values, the observed values are the same, but the missing values are filled in with a distribution of imputations that reflect the uncertainty about the missing data (Honaker & King, 2010). Auxiliary variables[13] included students' 2008 – 2009 and 2009 – 2010 state achievement test scores in all available subject areas (i.e., writing, social studies, and science for both math and reading models; reading for math models and math for reading models), as well as student BAM scores in models imputing SAT-9OE data and SAT-9OE scores in models imputing BAM data. Teacher scores on the MET Content Knowledge for Teaching assessment were used as auxiliary variables for the teacher years of experience in district and possession of advanced degree variables. Within the reading sample, 1.8% of student demographic data, 10.4% of state reading achievement test scores, 16.0% of SAT-9OE scores, 24.6% of data on teacher possession of a master's degree or higher, and 27.2% of data on years of experience teaching in the district were imputed. Within the math sample, 2.0% of student demographic data, 10.4% of state mathematics achievement test scores, 18.4% of BAM scores, 22.9% of

Table 3 Fully Unconditional Models of Assessment Scores

	SAT-9OE	State ELA	BAM	State math
Intercept (SE)	−0.006 (0.015)	−0.011 (0.018)	−0.008 (0.018)	−0.013 (0.020)
Between-student variance (σ^2)	0.842	0.756	0.768	0.718
Between-teacher variance (τ)	0.157	0.248	0.231	0.286
Total variance	0.999	1.003	0.999	1.004
Proportion of variance between students within schools[a]	0.843	0.753	0.769	0.715
Proportion of variance between teacher (intraclass correlation [ICC])[b]	0.157	0.247	0.231	0.285
Reliability (λ)	0.827	0.891	0.888	0.913

Note. ELA = English language arts; BAM = Balanced Assessment of Mathematics.
Source: Author's calculations using Measures of Effective Teaching (MET) data.
[a] $1 - (\tau/(\tau + \sigma^2))$. [b] $\text{ICC} = \tau/(\tau + \sigma^2)$.

data on teacher possession of a master's degree or higher, and 27.3% of data on years of experience teaching in the district were imputed.[14]

Results

The results of the unconditional models established the need for multilevel modeling to address variance at each conceptual level of the nested models. The first set of unconditional models indicated that a smaller proportion of the variance in SAT-9OE than in state reading achievement test scores was explained by differences between teachers and their teaching (16% and 25%, respectively), whereas the remainder was between students (see Table 3). A smaller proportion of the variance in BAM scores than in state mathematics achievement test scores lies between teachers (23% vs. 29%).

Next, I present the results of the multilevel models predicting achievement gaps on the MET constructed-response assessments. I follow this with a comparison of the size of achievement gaps estimated using SAT-9OE or BAM data with achievement gaps estimated using state reading or mathematics test data, respectively. Note that I focus specifically on Black – White and Hispanic – White achievement gaps because Black and Hispanic students have a history of underperforming on standardized tests in comparison to their White peers, whereas Asian students tend to perform at the same level or above White students on such tests.[15]

Achievement Gaps on MET Constructed-Response Reading and Mathematics Assessments

SAT-9OE Racial Achievement Gaps

In the first model of SAT-9OE scores, Black and Hispanic students were predicted to score significantly lower than White students (0.36 and 0.20 *SD*, respectively; $p < .001$; see Table 4) on the SAT-9OE without any controls for student or teacher characteristics. These estimates of Black – White and Hispanic – White achievement gaps in SAT-9OE scores were reduced only very slightly (to 0.34 and 0.17 *SD*, respectively; $p < .001$) by the addition of the measure of free and reduced-price lunch in Model 3. The estimated Black – White gap was diminished by half (to 0.16 *SD*s; $p < .001$) after the addition of prior reading achievement test scores to the model, and the estimated Hispanic – White gap was also significantly reduced, and no longer significant, after this addition (see Model 4). The addition of other student characteristics, in Model 5, and teacher and classroom characteristics, in Models 6 and 7, made virtually no difference in the size of estimated Black – White and Hispanic – White achievement gaps on the SAT-9OE.

SAT-9OE Income Achievement Gaps

As shown in Model 2 in Table 4, the income achievement gap in SAT-9OE scores, identified using the free and reduced-price lunch measure, was estimated at 0.12 *SD* without any control variables ($p < .001$). Model 3 shows that it was only slightly smaller (0.09 *SD*; $p < .001$) after adjusting for student race. After the addition of students' prior state reading achievement test scores in Model 4, the estimated income achievement gap was almost nonexistent (0.02 *SD*) and no longer statistically significant, indicating that race and, even more so, prior reading achievement accounted for the estimated income achievement gap in SAT-9OE scores among MET fourth- through eighth-grade students.

Table 4 Racial and Income Gaps in Student Scores on the SAT-9 Open-Ended Reading Assessment

Fixed effects	Model 1 Student race	Model 2 Student FRL	Model 3 Student race and FRL	Model 4 Student race, FRL, and prior achievement	Model 5 All student characteristics	Model 6 Student and teacher characteristics	Model 6 Student and teacher characteristics
Student-level equation							
Intercept	0.185***	0.055***	0.218***	0.073***	0.212***	0.198***	0.184***
Black[a]	−0.358***		−0.342***	−0.163***	−0.177***	−0.176***	−0.163***
Hispanic	−0.202***		−0.171***	−0.034	−0.015	−0.014	−0.006
Asian	0.097***		0.113***	0.139***	0.140***	0.141***	0.141***
Other race	−0.049		−0.031	0.026	0.016	0.016	0.022
Free and reduced-price lunch student		−0.117***	−0.094***	−0.021	−0.010	−0.010	−0.007
State ELA achievement test 2008–09				0.418***	0.373***	0.373***	0.364***
Grade					−0.013	−0.013	−0.012
Male					−0.247***	−0.247***	−0.247***
Gifted student					0.131***	0.132***	0.118***
Special education student					−0.328***	−0.328***	−0.326***
English language learner					−0.093***	−0.092***	−0.086***
Teacher-level equation							
Years of experience teaching in district						0.002	0.002
Master's or higher degree						0.010	0.013
Class 2008–09 state ELA test score average							0.077***
Random effects: variance components							
Intercept	0.138***	0.150***	0.133***	0.069***	0.066***	0.066***	0.062
Level 1	0.828	0.840	0.8288	0.707	0.682	0.682	0.682

Note: Table presents hierarchical linear modeling (HLM) coefficients: Outcome variables are standardized student test scores on the SAT-9 Open-Ended Reading Assessment. Free and reduced-price lunch eligibility is a proxy for a low-income family.

[a]Indicators of student race; comparison is to White students.

$^{*}p < .05$.

BAM Racial Achievement Gaps

In the initial model predicting BAM scores, Black and Hispanic students were predicted to score significantly lower than White students on the BAM (0.56 and 0.35 *SD*, respectively; $p < .001$; see Table 5). As was the case for the SAT-9OE, the estimated Black–White and Hispanic–White achievement gaps in BAM scores were reduced only slightly by the addition of the measure of free and reduced-price lunch (Model 3)—to 0.53 *SD* and 0.29 *SD* ($p < .001$), respectively. The estimated Black–White and Hispanic–White achievement gaps were reduced by more than half by the addition of state mathematics achievement test scores, to 0.25 *SD* and 0.13 *SD*, respectively, in Model 4 ($p < .001$). As was the case for the SAT-9OE, the addition of other student characteristics, in Model 5, and teacher and classroom characteristics, in Models 6 and 7, made virtually no difference in the size of Black–White and Hispanic–White achievement gap estimates on the BAM. Overall, results revealed that prior achievement accounted for more than half of the Black–White and Hispanic–White achievement gap estimates, whereas family income, as measured by free and reduced-price lunch eligibility and average prior achievement, a measure of average entering student ability for students taught by each teacher, accounted for far smaller portions of these estimated gaps. A small, but statistically significant, Black–White gap estimate and an even smaller, statistically significant Hispanic–White gap estimate remained, even after holding constant available student, teacher, and classroom characteristics.

BAM Income Achievement Gaps

The income achievement gap in BAM scores, as measured by eligibility for free and reduced-price lunches, was estimated at 0.22 *SD* without any control variables and 0.17 *SD* with controls for student race ($p < .001$; see Table 5). The estimated

Table 5 Racial and Income Gaps in Student Scores on the Balanced Assessment of Mathematics (BAM)

Fixed effects	Model 1 Student race	Model 2 Student FRL	Model 3 Student race and FRL	Model 4 Student race, FRL, and prior achievement	Model 5 All student characteristics	Model 6 Student and teacher characteristics	Model 7 Student and teacher and classroom characteristics
Student-level equation							
Intercept	0.307***	0.103***	0.356***	0.151***	0.167***	0.153***	0.140***
Black[a]	−0.562***		−0.526***	−0.247***	−0.247***	−0.252***	−0.240***
Hispanic	−0.348***		−0.285***	−0.130***	−0.106***	−0.107***	−0.101***
Asian	0.112***		0.147***	0.047*	0.060**	0.059**	0.061**
Other race	−0.220***		−0.203***	−0.073**	−0.071**	−0.072**	−0.065*
Free and reduced-price lunch student[b]		−0.221***	−0.165***	−0.038**	−0.024	−0.021	−0.017
State math achievement test 2008–09				0.606***	0.575***	0.575***	0.567
Grade					−0.011	−0.007	−0.006
Male					−0.061***	−0.061***	−0.060***
Gifted student					0.288***	0.287***	0.274***
Special education student					−0.202***	−0.202***	−0.200***
English language learner					−0.062***	−0.063***	−0.058***
Teacher-level equation							
Years of experience teaching in district						−0.003	−0.003*
Master's or higher degree						0.041*	0.047*
Class 2008–2009 state math test score average							0.066***
Random effects: variance components							
Intercept	0.184***	0.211***	0.173***	0.055***	0.052***	0.052***	0.048***
Level 1	0.736	0.761	0.733	0.484	0.476	0.476	0.476

Note: Table presents hierarchical linear modeling (HLM) coefficients: Outcome variables are standardized student test scores on the BAM mathematics assessment.

[a]Indicators of student race; comparison is to White students. [b]Free and reduced price lunch eligibility is a proxy for a low-income family. *$p < .05$. **$p < .01$. ***$p < .001$.

BAM income achievement gap was smaller but still significant after the addition of students' prior year state mathematics achievement test scores. In contrast to the SAT-9OE, the estimated BAM income achievement gap was not accounted for by the addition of prior achievement. However, the addition of other student characteristics, including gender and eligibility for gifted, English language learner, and special educational services, did render the estimated BAM income achievement gap almost nonexistent (0.02 *SD*) and statistically nonsignificant. Thus, this estimated gap was accounted for by student demographic characteristics and, even more so, by prior academic skills.

Comparisons of Achievement Gaps on MET Teaching Constructed-Response and State Assessments

To enable a comparison of racial and income achievement gaps on the MET constructed-response assessments to gaps on state reading achievement tests, I specified two additional sets of multilevel models, one for reading and another for mathematics, with state achievement test scores as the outcome variables (see Tables A1 and A2 in the appendix). Below, I present a comparison of the results for each test format, first for reading and subsequently for mathematics. All achievement gaps discussed in this section are statistically significant at the $p < .01$ level and lower.

SAT-9OE Reading Test Versus State Reading Tests

Prior to the addition of control variables, Black–White, Hispanic–White, and income achievement gaps were slightly smaller when measured by SAT-9OE data than when estimated using state reading achievement test data (see Figure 1). In the final models accounting for student, teacher, and classroom characteristics, all remaining achievement gap estimates were small, ranging from 0.04 *SD* to 0.16 *SD*. The estimated Black–White reading achievement gap was slightly larger

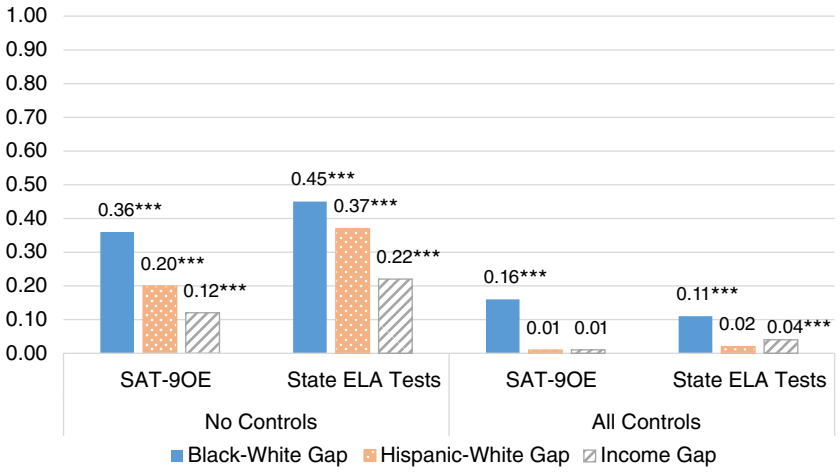

Figure 1 Estimates of racial and income achievement gaps using SAT-9OE scores and state reading test scores (*SD*); $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

when measured by the SAT-9OE than by the state reading tests (0.16 *SD* vs. 0.11 *SD*, respectively; $p < .001$). In contrast, the estimated Hispanic–White achievement gap was nearly nonexistent on both tests (0.01 *SD* vs. 0.02 *SD*, respectively) and was no longer significant for either test after the addition of all control variables, indicating that prior student academic achievement and average within-classroom student reading ability accounted for the differential reading test performance of Hispanic and White students on both the SAT-9OE and the state reading achievement tests. When estimated using SAT-9OE data, the income achievement gap in reading scores was almost zero (0.01 *SD*) and not significant after the addition of control variables. The estimated income achievement gap remained statistically significant, but very small and practically insignificant (0.04 *SD*; $p < .001$) when measured by the state reading assessment.

BAM Versus State Mathematics Tests

Prior to the addition of control variables to account for differences in student, teacher, and classroom characteristics, all three types of achievement gaps were similarly sized when measured using BAM data or scores from the state mathematics achievement tests used by MET districts, differing only by 0.03–0.05 *SD* (see Figure 2). In the final models controlling for student, teacher, and classroom characteristics, estimated racial achievement gaps were larger when using data from the BAM than when using data from the state mathematics tests. The estimated Black–White achievement gap was twice as large on the BAM as on state mathematics tests (0.24 *SD* and 0.12 *SD*, respectively; $p < .001$), and the estimated Hispanic–White achievement gap was three times as large on the BAM (0.10 *SD*; $p < .001$) as on the state mathematics tests (0.03 *SD*; $p < .01$) in the final models. Whereas the estimated income achievement gap in the BAM models was accounted for by student, teacher, and classroom characteristics and was no longer significant at 0.02 *SD*, a small, statistically significant gap (0.05 *SD*; $p < .001$) remained in the final model predicting state mathematics test scores. As was the case for the reading achievement gaps in models with all controls, the remaining income achievement gap in math scores was so small as to be practically insignificant.[16]

Discussion

The MET data provide a useful opportunity to explore the extent to which racial and income achievement gaps may vary by test item format. Using SAT-9OE and BAM test scores and 2009–2010 state reading and mathematics test scores for the over 25,000 fourth- through eighth-grade students in the six-state MET study, I found that estimated racial and income achievement gaps vary by test item format and cognitive demand. The finding that some racial achievement gaps may be slightly larger on constructed-response assessments, after controlling for other factors, aligns with those of some prior studies of racial and ethnic score differences by item format (Bond, 1995; Dimitrov, 1999; Feinberg, 1990; Linn et al., 1991; Shannon & Bylsma, 2002). At the same time, the finding that reading achievement gaps are smaller on constructed-response assessments, before controlling for any other factors, aligns with those of other studies (Arthur et al., 2002;

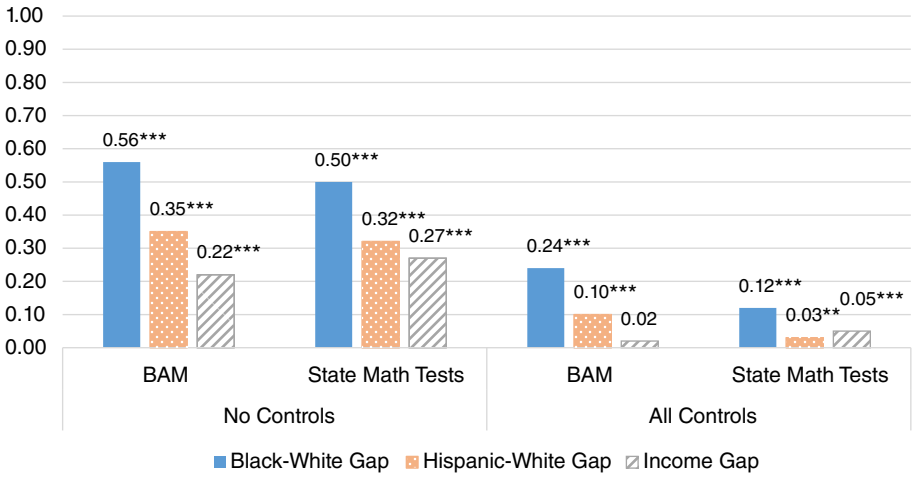

Figure 2 Estimates of racial and income achievement gaps using BAM scores and state mathematics test scores (*SD*); $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Badger, 1995; O'Neil & Brown, 1998). The findings also contradict those of a prior study of socioeconomic differences in test performance by test item format (Wright et al., 2016) in that my findings indicate that income achievement gaps may be slightly smaller on constructed-response assessments.

Importantly, based on the higher proportion of variance in state achievement test scores attributed to teachers, the study results also indicate that teachers and their practices may matter slightly less when it comes to scores on more cognitively demanding constructed-response assessments. This may simply be because there was no "teaching to the test" prior to the administrations of the SAT-9OE and BAM assessments. It may also indicate that constructed-response assessments, in particular those that are more cognitively demanding, are less sensitive to differences in teaching. Given that Grossman et al. (2014), using MET data, found that the SAT-9OE in particular may be more sensitive to cognitively demanding teaching, perhaps the amount of cognitively demanding teaching is what varies across MET study classrooms.

Reading

The results of the multilevel models of reading test scores show that reading achievement gaps may vary by cognitive demand and item type. Although Black and Hispanic students may actually underperform White students by less on a constructed-response reading assessment relative to a multiple choice assessment, differences in prior reading achievement account for this smaller achievement gap. Models predicting achievement on the SAT-9OE reading assessment revealed smaller Black–White, Hispanic–White, and income achievement gap estimates than models predicting state reading test scores, without adjusting for the influence of student, teacher, and classroom characteristics on student reading achievement.

However, it is important to note that results for reading models were mixed; achievement gaps were smaller on the SAT-9OE than state reading tests *before* models controlled for other factors influencing the size of gaps. Thus, before accounting for prior reading achievement, family income, and status as a gifted or special education student or an English language learner, the estimated racial and income achievement gaps were larger on state reading tests, even given their lower level of cognitive demand. Once student, teacher, and classroom characteristics were included in the model, the remaining achievement gap estimates were small or nonexistent, and the estimated Black–White achievement gap was slightly larger on the SAT-9OE than on the state reading assessment.

This was not the case for the Hispanic–White reading achievement gap; Hispanic students did not score significantly lower than White students after student, teacher, and classroom characteristics were accounted for in both the state reading test model and the SAT-9OE model. In fact, no significant Hispanic–White achievement remained on either test. The income reading achievement gap was also accounted for by control variables when estimated using the SAT-9OE. When estimated using the state reading tests, it was so small as to be practically nonsignificant. Thus, the Hispanic–White and income reading achievement gaps were explained away by the control variables, whereas an estimated Black–White

achievement gap remained on both assessments and was slightly larger on the SAT-9OE constructed-response assessment.

Mathematics

Like the reading results, the mathematics results indicate that achievement gaps may vary by the level of cognitive demand of the assessment, which in this study was intertwined with item type. In models not accounting for differences in student, teacher, and classroom characteristics, estimated Black – White and Hispanic – White mathematics achievement gaps were just slightly larger using data from the BAM constructed-response assessments, whereas the estimated income achievement gap was just slightly smaller using data from the BAM than using state mathematics test scores. Once student, teacher, and classroom characteristics were accounted for, leaving only small remaining gaps, the Black – White achievement gap was twice as large on the BAM as on the state mathematics tests, and the Hispanic – White achievement gap was three times as large on the BAM as on the state mathematics tests. Thus, estimated racial achievement gaps may be larger on a constructed-response mathematics assessment selected for its higher level of cognitive demand.

In contrast, the income achievement gap was smaller on the BAM than on the state mathematics tests, both before and after accounting for student, teacher, and classroom factors that may influence the size of achievement gaps. In fact, in the final model accounting for these factors, the income achievement gap estimated using BAM data was no longer significant. This means that the remaining racial achievement gaps were larger on the BAM constructed-response assessment than on largely multiple-choice state assessments, and the reverse was true for income achievement gaps; however, the very small size of the remaining income achievement gap (0.05 *SD*) translates to little to no practical significance in terms of differences in mathematics achievement between lower and higher income students.

As was the case for reading, the results of this study highlight the possibility of greater racial achievement gaps on more cognitively demanding mathematics assessments, which indicates a potential for achievement gaps to differ across various types of cognitive skills or on assessments using different types of items. In addition to highlighting the existence of variations in achievement gaps by assessment item type and level of cognitive demand, the results of this study provide several insights into the relative contribution of various factors to reading and mathematics achievement gaps.

Factors Explaining Reading and Mathematics Achievement Gaps

Although eligibility for free and reduced-price lunch reduced the size of estimated Black – White and Hispanic – White achievement gaps, and race and ethnicity diminished estimated income achievement gaps on MET constructed-response assessments and on the state assessments, prior subject-specific achievement reduced these gaps by a much greater amount on both types of assessments. Prior achievement even rendered the Hispanic – White reading achievement gap and the income achievement gap nonsignificant using data from the SAT-9OE and considerably diminished estimates of both gaps in models using data from the BAM or state reading and mathematics tests. Prior achievement is also reflective of early achievement gaps; research documents that these gaps begin in the earliest years of life and are influenced by children's experiences both in and out of school (Brooks-Gunn & Duncan, 1997; Hart & Risley, 1995).

Another potential contributor to reading and mathematics achievement gaps is the mix of peer abilities to which children are exposed in their classrooms. Average within-classroom student ability rendered the Hispanic – White reading achievement gap insignificant on the state reading test, reduced the size of all three types of gaps on the state reading and mathematics tests, and reduced the size of racial achievement gaps on the BAM. This finding aligns with prior research on the importance of peer academic abilities (Epple & Romano, 2011).

Limitations

The study limitations include the fact that MET researchers used opportunity sampling and, thus, the data are not nationally representative or representative of specific school districts (White et al., 2014). The study combines scores from six different state tests to provide measures of state reading and mathematics test scores, and the standardization of scores within grades and then districts accounts for variations in test score scales. However, a remaining issue could be variations in assessment content due to differences in learning standards between the MET states, particularly prior to the adoption of the Common Core State Standards. Moreover, a content study of the state tests and BAM and SAT-9OE assessments

has not been conducted. Nonetheless, prior to selecting the BAM and SAT-9OE as more cognitively demanding assessments than state achievement tests at the time, MET researchers did conduct an alignment study to verify that the content of these tests aligned with the state standards for each of the six participating districts. Thus, differences by content are assumed to be minimal, although some differences likely exist given variations in state learning standards at the time of the MET study. Unfortunately, MET data users are not permitted to make comparisons between states, because the data are not representative of MET districts or states, and thus the possibility of between-state differences could not be explored in this study.

Another remaining issue is that the study does not address school-level contextual effects, although the estimates were adjusted for the classroom-level contextual effect of average peer academic achievement in reading or mathematics, respectively. Even so, bias due to classroom- and school-level endogeneity, or the confounding of student covariates with unobserved classroom or school characteristics, may still be an issue (Castellano, Rabe-Hesketh, & Skrondal, 2014).

Yet racial and income achievement gaps remain a clear and useful indicator of inequalities in educational opportunities within and between schools and localities. The accurate measurement of these gaps could help to better inform potential solutions. The study provides observational evidence on the relevance of test item format when measuring achievement gaps and highlights a possible need to take a more fine-grained approach to measuring achievement gaps inclusive of a broad range of cognitive skills.

Summary and Implications

In summary, this study uses large-scale data from constructed-response assessments and 2009–2010 state assessments to clarify that racial and income achievement gaps may appear smaller or larger depending on the type of assessment items used. In contrast with data from largely multiple-choice state assessments, data from the more cognitively demanding MET constructed-response assessments produce smaller estimates of achievement gaps in reading before and a larger estimate of the Black–White reading gap after models are adjusted for explanatory factors. Estimated Black–White and Hispanic–White mathematics achievement gaps were slightly larger using BAM data, whereas the estimated income achievement gap was slightly smaller using BAM data. These differing achievement gaps between assessments of different item types and cognitive demand could indicate a need for instructional shifts toward greater support for higher order thinking skills for some students from disadvantaged backgrounds. Prior research has shown that teachers may not equally support the higher order thinking skills of lower achieving students (Raudenbush, Rowan, & Cheong, 1993), who are more often from disadvantaged backgrounds. This may be one way achievement gaps are inadvertently perpetuated, and it should be a focus of future research.

The study findings also indicate a need for further research into whether achievement gaps may be best measured using a mix of items of different types and different levels of cognitive demand, providing the greatest potential to measure a broad range of cognitive skills. Whereas some tests used to identify achievement gaps, such as the NAEP (Hemphill & Vanneman, 2011), may already meet these criteria, many state and local assessments may not. One promising recent development is the increased focus within the field of education on standardized assessments that measure a range of thinking skills, including higher order skills, due to the curricular shifts called for by the Common Core State Standards (Educational Testing Service, 2016). The greater cognitive demand of some more recent assessments may better measure achievement gaps by accounting for a broader range of cognitive abilities. Moreover, the curricular shifts called for by the new standards may be influencing the extent to which students of all backgrounds are exposed to teaching that supports higher order thinking skills.

This study also provides observational evidence on the importance of teacher effectiveness previously demonstrated by a multitude of studies (Hanushek, 2002; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). The finding that a substantial proportion of the variance in both types of test scores (constructed-response and 2009–2010 state assessments) lies between teachers suggests that differences in instruction in particular may contribute to variations in student achievement. A greater proportion of variance in the unconditional models was attributed to teacher effects on the BAM and the state mathematics and reading tests, highlighting the challenges and possibilities of teaching a range of cognitive skills in both subject areas.

Increasing the effectiveness of teaching, and with it school quality, remains an important pathway to improve educational outcomes and, accordingly, to address achievement gaps (Reardon, 2017). Some evidence indicates that teachers

who teach in ways that promote critical thinking and reasoning are better at raising achievement on more cognitively demanding assessments, such as the BAM and the SAT-9OE (Grossman et al., 2014). Moreover, constructed-response assessments such as the BAM and the SAT-9OE may even foster conceptual understanding and increase use of higher order thinking skills among students (Jensen et al., 2014; Scully, 2017). One potential avenue for future research, based on the findings of this observational study, would be using longitudinal, nationally representative data to explore the extent to which achievement gaps might vary across different types of cognitive skills, parsing out differences due to item type from those due to cognitive demand. Further research might then aim to identify the various teaching approaches that would best support both lower and higher order thinking skills.

Notes

1 Under NSLP policy, students whose household income is less than 130% of the poverty line qualify for free lunch, and students whose household income is between 130% and 185% of the poverty line qualify for reduced-price lunch (Domina et al., 2017; U.S. Department of Agriculture Food and Nutrition Service, 2018).
2 Note that data for the Reardon study were drawn from multiple nationally representative data sets.
3 Note that the SAT-9OE and BAM were administered only to students in Grades 4 – 8; thus students in Grade 9 are not included in this study.
4 Although I do not model students within schools in this study, I note here that the ELA sample included data from 215 schools and the mathematics sample included data from 201 schools. The distribution of students was not proportional across districts; districts varied in size and study agreements stipulated that entire districts agree to participate or not.
5 This includes BAM and SAT-9OE data, as well as student survey data. State achievement test scores were provided to MET researchers in district administrative data files; thus multiple years of prior state achievement test scores are available. In this study I took advantage of this by including test scores from the prior year as control variables.
6 Note that MET researchers did not use the full SAT-9 assessment; rather, they modified it for their study by using only the SAT-9OE and no other portions of the full SAT-9 assessment (D. McCaffrey, personal communication, June 1, 2019).
7 Pearson purchased the publisher of the SAT-9OE, Harcourt, prior to the MET study.
8 Technical documentation was obtained from three of the six MET states for the 2009 – 2010 state reading and math assessments; for the three remaining states information on item formats and number of items was obtained from the 2009 NAEP state mapping studies of 2008 – 2009 tests, which were not conducted on 2009 – 2010 tests.
9 Note that the MET study was conducted prior to the majority of states' adoption of the Common Core State Standards; thus, at the time of the study, learning standards and assessment content may have differed across the six MET states.
10 In 2010, 48% of US public school students qualified for free or reduced-price lunch (National Center for Education Statistics, 2012), a frequently used proxy for low-income status, as noted previously.
11 This variable represents the total years the teacher taught in the district at the time of the MET data collection. The total years of teaching variable was missing much more data than the total years of teaching in the district variable (e.g., 63% for the overall years of teaching variable vs. 27.2% for years of experience teaching in the district, for the ELA sample). The decision was made to use years of experience teaching in the district to minimize the amount of data requiring imputation. In the majority of cases with data, teachers had spent the majority of their teaching years in the district they were in at the time of the study.
12 I used the multiple imputation module of SPSS Version 24, which uses a chained equation routine and produces five imputed data sets for analysis. Constraints were included to ensure consistent variance before and after imputation.
13 In most cases, it can be assumed that data are not missing completely at random but instead are missing at random (MAR); that is, missing data on particular measures are conditionally independent in the fully specified model (Rubin, 1987). Auxiliary variables are chosen because they are correlated with variables with missing data to support the MAR assumption.
14 Multiple imputation is sound even for large percentages of missing values because the application of Rubin's rules preserves confidence intervals and Type I error rates (Rubin, 1987). The amount of missing data is not as important as the convergence of imputation models (Dong & Peng, 2013).
15 Although achievement gaps may exist among Asian subgroups, this study does not focus on these subgroup differences. Similarly, differences may exist among Hispanic or Latino subgroups, but this study does not focus on these differences.
16 To explore the extent to which racial achievement gaps varied by family income and vice versa, I specified additional models with interaction terms representing Black, Hispanic, Asian, and other race students eligible for FRL. Free and reduced-priced lunch students from some minority groups were predicted to score slightly lower on achievement tests than White FRL students, and minority students not eligible for FRL were sometimes still predicted to score lower than White students ineligible for FRL. The extent of these effects varied by the type of assessment and were consistently small, only less than 0.01 *SD* in models with all control variables, if they were even significant.

References

Ackerman, B. P., & Brown, E. D. (2010). Physical and psychosocial turmoil in the home and cognitive development. In *Chaos and its influence on children's development: An ecological perspective* (pp. 35–47). Washington, DC: American Psychological Association.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Arthur, W., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed-response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology*, *55*(4), 985–1008.

Badger, E. (1995). The effects of expectations on achieving equity in state-wide testing: Lessons from Massachusetts. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 289–308). Boston, MA: Kluwer Academic.

Balanced Assessment in Mathematics Project. (1995). *Assessing mathematical understanding and skills effectively*. Retrieved from https://hgse.balancedassessment.org/amuse.html

Barton, P. E., & Coley, R. J. (2009). *Parsing the achievement gap II*. Princeton, NJ: Educational Testing Service.

Barton, P. E., & Coley, R. J. (2010). *The Black–White achievement gap: When progress stopped*. Princeton, NJ: Educational Testing Service.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice*, *14*(4), 21–24. https://doi.org/10.1111/j.1745-3992.1995.tb00885.x

Brooks-Gunn, J., & Duncan, G. J. (1997). The effects of poverty on children. *The Future of Children*, *7*(2), 55–71. https://doi.org/10.2307/1602387

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q., & Luppescu, S. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics*, *39*(5), 333–367. https://doi.org/10.3102/1076998614547576

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, A. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: National Center for Education Statistics.

Darling-Hammond, L. (2014). Closing the achievement gap: A systemic view. In J. V. Clark (Ed.), *Closing the achievement gap from an international perspective*. Berlin, Germany: Springer Science and Business Media.

Darling-Hammond, L. (2017). *Developing and measuring higher order skills: Models for state performance assessment systems*. Washington, DC: Council of Chief State School Officers.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55–77. https://doi.org/10.1207/s15324818ame1301_3

Dimitrov, D. M. (1999, April). *Mathematics and science achievement profiles by gender, race, ability, and type of item response*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, QC.

Domina, T., Brummet, Q., Pharris-Ciurej, N., Porter, S. R., Penner, A., Penner, E., & Sanabria, T. (2017). *Capturing more than poverty: School free and reduced-price lunch data and household income* [Working paper]. Retrieved from the Center for Administrative Records Research and Applications website https://www.census.gov/content/dam/Census/library/working-papers/2017/adrm/carra-wp-2017-09.pdf

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222. https://doi.org/10.1186/2193-1801-2-222

Duncan, G. J., & Brooks-Gunn, J. (1997). *Consequences of growing up poor*. New York, NY: Russell Sage Foundation.

Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 47–70). New York, NY: Russell Sage Foundation.

Educational Testing Service. (2016). *The road ahead for state assessments: What the assessment consortia built, why it matters, and emerging options*. Retrieved from https://www.ets.org/s/k12/pdf/coming_together_the_road_ahead.pdf

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Epple, D., & Romano, R. E. (2011). Peer effects in education: A survey of the theory and evidence. In J. Benhabib, A. Bisin, & M. O. Jackson (Eds.), *Handbook of social economics* (Vol. 1, pp. 1053–1163). New York, NY: Elsevier.

Farkas, G. (2011). Middle and high school skills, behaviors, attitudes, and curriculum enrollment, and their consequences. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 71–90). New York, NY: Russell Sage Foundation.

Feinberg, L. (1990). Multiple-choice and its critics. *The College Board Review*, *157*, 12–17.

Ferguson, R. F., & Hirsch, E. (2015). How working conditions predict teaching quality and student outcomes. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 332–380). San Francisco, CA: Jossey-Bass.

Fryer, R. G., & Levitt, S. D. (2004). Understanding the Black–White test score gap in the first two years of school. *Review of Economics and Statistics*, *86*(2), 447–464. https://doi.org/10.1162/003465304323031049

García, E., & Weiss, E. (2017). *Reducing and averting achievement gaps*. Retrieved from the Economic Policy Institute website https://www.epi.org/publication/reducing-and-averting-achievement-gaps/

Grodsky, E., & Pager, D. (2001). The structure of disadvantage: Individual and occupational determinants of the Black–White wage gap. *American Sociological Review*, *66*(4), 542–567. https://doi.org/10.2307/3088922

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *43*(6), 293–303. https://doi.org/10.3102/0013189X14544542

Hansen, M., Levesque, E., Quintero, D., & Valant, J. (2018). *Have we made progress on achievement gaps? Looking at evidence from the new NAEP results*. Retrieved from the Brookings Institute website https://www.brookings.edu/blog/brown-center-chalkboard/2018/04/17/have-we-made-progress-on-achievement-gaps-looking-at-evidence-from-the-new-naep-results/

Hanushek, E. A. (2002). Teacher quality. In L. T. Izumi & W. M. Evers (Eds.), *Teacher quality* (pp. 1–12). Stanford, CA: Hoover Press.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test series: Ninth edition. Technical Data Report*. San Antonio, TX: Author.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H Brookes Publishing.

Hemphill, F. C., & Vanneman, A. (2011). *Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics.

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561–581. https://doi.org/10.1111/j.1540-5907.2010.00447.x

Inverness Research. (2014). *What teachers say about the benefits of MAP classroom challenges*. Retrieved from http://inverness-research.org/mars_map/reports/2014-02_Rpt-MAP_WhatTeacherSay-Benefits.pdf

Jencks, C., & Phillips, M. (1998). *The Black–White test score gap*. Washington, DC: Brookings Institution Press.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test … or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, *26*(2), 307–329. https://doi.org/10.1007/s10648-013-9248-9

Jones, L. V. (1984). White–Black achievement differences: The narrowing gap. *American Psychologist*, *39*(11), 1207–1213. https://doi.org/10.1037/0003-066X.39.11.1207

Kaiser Family Foundation. (2017). *Poverty rate by race/ethnicity*. Retrieved from https://www.kff.org/other/state-indicator/poverty-rate-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

Kane, T. J., & Cantrell, S. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.

Kirk, D. S., & Sampson, R. J. (2011). Crime and the production of safe schools. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 397–418). New York, NY: Russell Sage Foundation.

Lareau, A. (2011). *Unequal childhoods: Class, race, and family life*. Oakland: University of California Press.

Lawrence, I. M., Lyu, C. F., & Feigenbaum, M. D. (1995). *DIF data on free-response SAT I mathematical items*. Princeton, NJ: Educational Testing Service.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.

Legal Information Institute. (n.d.). *Brown v. Board of Education*. Retrieved from https://www.law.cornell.edu/supremecourt/text/347/483

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21. https://doi.org/10.3102/0013189X020008015

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current state and directions for next-generation assessment* (Research Report No. RR-14-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12009

Livingston, S. A. (2009). *Constructed-response test questions: Why we use them; how we score them*. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections11.pdf

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234–250. https://doi.org/10.1111/j.1745-3984.1994.tb00445.x

Mackey, A. P., Finn, A. S., Leonard, J. A., Jacoby-Senghor, D. S., West, M. R., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2015). Neuroanatomical correlates of the income-achievement gap. *Psychological Science*, *26*(6), 925–933. https://doi.org/10.1177/0956797615572233

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan*, *80*(9), 688–695.

Mathematics Assessment Collaborative. (2003). *MAC report on the 2003 tests*. Morgan Hill, CA: Silicon Valley Mathematics Initiative.

Mathematics Assessment Collaborative. (2004). *MAC report on the 2004 tests*. Morgan Hill, CA: Silicon Valley Mathematics Initiative.

Mathematics Assessment Collaborative. (2005). *MAC report on the 2005 tests*. Morgan Hill, CA: Silicon Valley Mathematics Initiative.

Mathematics Assessment Project. (2015). *Assessing 21st century math: About the Math Assessment Project*. Retrieved from https://www.map.mathshell.org/background.php

Murnane, R. J., & Levy, F. (1996). *Teaching the new basic skills: Principles for educating children to thrive in a changing economy*. New York, NY: Free Press.

National Assessment Governing Board. (2017a). *Mathematics framework for the 2017 National Assessment of Educational Progress*. Retrieved from https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/mathematics/2017-math-framework.pdf

National Assessment Governing Board. (2017b). *Reading framework for the 2017 National Assessment of Academic Progress*. Retrieved from https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/reading/2017-reading-framework.pdf

National Assessment of Educational Progress. (2019). *Sample questions grade 8 mathematics reading science: General information about The Nation's Report Card*. Washington, DC: National Center for Education Statistics.

National Center for Education Statistics. (2011a). *A profile of state assessment standards: 2009*. Retrieved from https://nces.ed.gov/nationsreportcard/studies/statemapping/profile_standards_2009.aspx

National Center for Education Statistics. (2011b). *Students meeting state proficiency standards and performing at or above the NAEP proficient level: 2009*. Retrieved from https://nces.ed.gov/nationsreportcard/studies/statemapping/2009_naep_state_table.aspx

National Center for Education Statistics. (2018). *Achievement gaps*. Retrieved from https://nces.ed.gov/nationsreportcard/studies/gaps/

North Carolina Department of Education. (2010a). *North Carolina statewide testing program raw scores by achievement level end-of-grade mathematics, edition 3, 2009–2010*. Raleigh, NC: North Carolina State Department of Education.

North Carolina Department of Education. (2010b). *North Carolina statewide testing program raw scores by achievement level end-of-grade reading, edition 3, 2009–2010*. Raleigh, NC: North Carolina State Department of Education.

O'Neil, H. F., Jr., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education*, *11*(4), 331–351. https://doi.org/10.1207/s15324818ame1104_3

Organisation for Economic Co-operation and Development. (2016). *Country note: Key findings from PISA 2015 for the United States*. Retrieved from https://www.oecd.org/pisa/PISA-2015-United-States.pdf

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology*, *34*, 181–209. https://doi.org/10.1146/annurev.soc.33.040406.131740

Patz, R. J., Wilson, M., & Hoskens, M. (1997). *Optimal rating procedures and methodology for NAEP open-ended items*. Retrieved from ERIC database (ED417204) https://files.eric.ed.gov/fulltext/ED417204.pdf

Pearson Education. (2018). *Stanford Open-Ended Reading Assessment*. Retrieved from https://www.pearsonassessments.com/learningassessments/products/100000689/stanford-open-ended-reading-assessment.html

Phillips, M. (2011). Parenting, time use, and disparities in academic outcomes. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 207–228). New York, NY: Russell Sage Foundation.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*(4), 3–14. https://doi.org/10.1111/j.1745-3992.2010.00189.x

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Bryk, A., & Congdon, R. (2013). *HLM 7.01 for Windows* [Hierarchical linear and nonlinear modeling software]. Skokie, IL: Scientific Software International.

Raudenbush, S. W., Marshall, J., & Art, E. (2011). Year-by-year and cumulative impacts of attending a high-mobility elementary school on children's mathematics achievement in Chicago. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 359–376). New York, NY: Russell Sage Foundation.

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal*, *30*(3), 523–553. https://doi.org/10.3102/00028312030003523

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 91–116). New York, NY: Russell Sage Foundation.

Reardon, S. F. (2017). *Educational opportunity in early and middle childhood: Variation by place and age*. Retrieved from the Center for Education Policy analysis website https://cepa.stanford.edu/sites/default/files/wp17-12-v201803.pdf

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Researcher*, *47*(5), 284–294. https://doi.org/10.3102/0013189X18762105

Research for Action. (2015). *MDC's influence on teaching and learning*. Retrieved from http://8rri53pm0cs22jk3vvqna1ub-wpengine.netdna-ssl.com/wp-content/uploads/2015/10/MDC_Influence_on_Teaching_Learning_February_2015.pdf

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252. https://doi.org/10.1257/0002828041302244

Rothstein, R. (2004). The achievement gap: A broader picture. *Educational Leadership*, *62*(3), 40–43.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545–571. https://doi.org/10.1207/s15327906mbr3304_5

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation*, *22*(4), 447–464.

Shannon, G. S., & Bylsma, P. (2002). *Addressing the achievement gap: A challenge for Washington State educators*. Retrieved from ERIC database (ED474392) https://files.eric.ed.gov/fulltext/ED474392.pdf

Sharkey, P. (2013). *Stuck in place: Urban neighborhoods and the end of progress toward racial equality*. Chicago, IL: University of Chicago Press.

Shepard, L. (1997). *Measuring achievement: What does it mean to test for robust understanding?* Retrieved from https://www.ets.org/Media/Research/pdf/PICANG3.pdf

Stanford University Center for Education Policy Analysis. (n.d.). *Racial and ethnic achievement gaps*. Retrieved from http://cepa.stanford.edu/educational-opportunity-monitoring-project/achievement-gaps/race/

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Synergis Education. (2018). *Bloom's Taxonomy and Webb's Depth of Knowledge*. Retrieved from https://www.synergiseducation.com/blooms-taxonomy-and-webbs-depth-of-knowledge/

Taylor, C. S., & Lee, Y. (2011). Ethnic DIF in reading tests with mixed item formats. *Educational Assessment*, *16*, 35–68. https://doi.org/10.1080/10627197.2011.552039

U.S. Department of Agriculture Food and Nutrition Service. (2009). *Child nutrition programs: Income eligibility guidelines*. Retrieved from https://www.govinfo.gov/content/pkg/FR-2009-03-27/pdf/E9-6806.pdf

U.S. Department of Agriculture Food and Nutrition Service. (2018). *School meals: Income eligibility guidelines*. Retrieved from https://www.fns.usda.gov/school-meals/income-eligibility-guidelines

Vanneman, A., Hamilton, L., Baldwin Anderson, J., & Rahman, T. (2009). *How Black and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics.

White, M., Rowan, B., Alter, G., & Greene, C. (2014). *User guide to the measures of effective teaching longitudinal database (MET LDB)*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE–Life Sciences Education*, *15*(2), 1–16. https://doi.org/10.1187/cbe.15-12-0246

Yuan, K., & Le, V.N. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests*. Retrieved from the RAND Corporation website https://doi.org/10.7249/WR967

Appendix A.

State Reading and Mathematics Achievement Tests in Measures of Effective Teaching Study States 2008–2009 or 2009–2010: Item Formats and Numbers of Items

	Reading				Mathematics			
	Multiple choice		Constructed response		Multiple choice		Constructed response	
State	*N*	%	*N*	%	*N*	%	*N*	%
Colorado (2009–2010)								
Grade 4	56	80%	14	20%	54	78%	15	22%
Grade 5	56	80%	14	20%	54	78%	15	22%
Grade 6	56	80%	14	20%	45	75%	15	25%
Grade 7	56	80%	14	20%	45	75%	15	25%
Grade 8	56	80%	14	20%	45	75%	15	25%
Average %	56	80%	14	20%	49	76%	15	24%
Florida (2009–2010)								
Grade 4	48	90%	6	10%	48	100%	0	0%
Grade 5	53	100%	0	0%	38	89%	7	11%
Grade 6	53	100%	0	0%	38	100%	0	0%
Grade 7	53	100%	0	0%	38	100%	0	0%
Grade 8	48	90%	6	10%	33	89%	7	11%
Average %	51	96%	2	4%	39	96%	3	4%
New York (2009–2010)								
Grade 4	28	65%	7	35%	30	62%	18	38%
Grade 5	24	75%	2	25%	26	76%	8	24%
Grade 6	26	62%	8	38%	25	71%	10	29%
Grade 7	30	73%	4	27%	30	79%	8	21%
Grade 8	26	62%	8	38%	27	60%	18	40%
Average %	27	67%	6	33%	28	70%	12	30%
North Carolina (2008–2009)								
Grade 4	58	100%	0	0%	82	100%	0	0%
Grade 8	62	100%	0	0%	80	100%	0	0%
Average %	60	100%	0	0%	81	100%	0	0%
Tennessee (2008–2009)								
Grade 4	67	100%	0	0%	67	100%	0	0%
Grade 8	67	100%	0	0%	67	100%	0	0%
Average %	67	100%	0	0%	67	100%	0	0%
Texas (2008–2009)								
Grade 4	40	100%	0	0%	41	100%	1	3%
Grade 8	48	100%	0	0%	49	100%	1	2%
Average %	44	100%	0	0%	45	100%	1	3%

Note. Data are from the following: Colorado Department of Education, (n.d.). Retrieved from (https://www.cde.state.co.us/assessment/coassess-additionalresources). *Florida Comprehensive Assessment Test Design Summary* by Florida Department of Education (2009). Retrieved from http://www.fldoe.org/core/fileparse.php/7490/urlt/fc05designsummary.pdf. National Assessment of Educational Progress (2011b). *Guide to the Grades 3–8 Testing Program in English Language Arts and Mathematics* by New York State Education Department (n.d.). Retrieved from http://www.p12.nysed.gov/assessment/ei/archive/gr3-8guide10.pdf. North Carolina Department of Education (2010a). North Carolina Department of Education (2010b).

Appendix B.

Results of Multilevel Models Predicting Achievement Gaps on State Achievement Tests Used in Measures of Effective Teaching Districts

Table B1 Racial and Income Gaps in Student Scores on State English Language Arts (ELA) Achievement Tests

Fixed effects	Model 1 Student race	Model 2 Student FRL	Model 3 Student race and FRL	Model 4 Student race, FRL, and prior achievement	Model 5 All student characteristics	Model 6 Student and teacher characteristics	Model 7 Teacher and classroom characteristics
Student-level equation							
Intercept	0.275***	0.104***	0.335***	0.103***	0.119***	0.115***	0.095***
Black[a]	−0.451***		−0.421***	−0.135***	−0.133***	−0.134***	−0.112***
Hispanic	−0.366***		−0.309***	−0.084***	−0.029*	−0.029*	−0.015
Asian	−0.037		−0.007	0.040*	0.065***	0.065***	0.066***
Other race	−0.180***		−0.147***	−0.061**	−0.053*	−0.053*	0.043
Free and reduced-price lunch student		−0.220***	−0.174***	−0.055***	−0.044***	−0.044***	−0.039***
Grade					−0.007	−0.007	−0.006
Male					−0.063***	−0.063***	−0.063***
Gifted student					0.270***	0.270***	0.247***
Special education student					−0.245***	−0.245***	−0.242***
English language learner					−0.163***	−0.163***	−0.153***
State ELA test score 2008–2009				0.709***	0.660***	0.660***	0.647***
Teacher-level equation							
Years of experience teaching in district						−0.001	−0.001
Master's or higher degree						0.004	0.006
Class 2008–2009 state ELA test score average							0.098***
Random effects: variance components							
Intercept	0.205***	0.230***	0.195***	0.029***	0.025***	0.025***	0.018***
Level 1	0.737	0.748	0.733	0.383	0.372	0.372	0.372

Note. FRL, free and reduced price lunch. Table presents hierarchical linear modeling (HLM) coefficients: Outcome variables are standardized student test scores on 2009–2010 state ELA assessments.

[a]Indicators of student race; comparison is to White students.

*$p < .05$. **$p < .01$. ***$p < .001$.

Table B2 Racial and Income Gaps in Student Scores on State Math Achievement Test

Fixed effects	Model 1 Student race	Model 2 Student FRL	Model 3 Student race and FRL	Model 4 Student race, FRL, and prior achievement	Model 5 All student characteristics	Model 6 Student and teacher characteristics	Model 7 Teacher and classroom characteristics
Student-level equation							
Intercept	0.266***	0.122***	0.332***	0.091***	0.092***	0.081***	0.069***
Black[a]	−0.499***		−0.450***	−0.122***	−0.123***	−0.126***	−0.115***
Hispanic	−0.321***		−0.237***	−0.052***	−0.039**	−0.040**	−0.033**
Asian	0.182***		0.229***	0.108***	0.116***	0.115***	0.117***
Other race	−0.203***		−0.180***	−0.026	−0.025	−0.026	−0.019
FRL student		−0.268***	−0.221***	−0.068***	−0.055***	−0.054***	−0.050***
Grade					−0.004	−0.002	0.000
Male					−0.009	−0.009	−0.009
Gifted student					0.224***	0.223***	0.211***
Special education student					−0.221***	−0.220***	−0.219***
English language learner					−0.045**	−0.045**	−0.041**
State math test score 2008–09				0.731***	0.704***	0.704***	0.697***
Teacher-level equation							
Years of experience teaching in district						−0.001	−0.002
Master's or higher degree						0.029	0.036 *
Class 2008–09 state math test score average							0.081***
Random effects: variance components							
Intercept	0.236***	0.259***	0.219***	0.052***	0.049***	0.049***	0.043***
Level 1	0.690	0.708	0.684	0.318	0.312	0.312	0.312

Note: Table presents hierarchical linear modeling (HLM) coefficients: Outcome variables are standardized student test scores on 2009–2010 state math assessments.

[a]Indicators of student race; comparison is to White students.

*$p < .05$. **$p < .01$. ***$p < .001$.

Suggested citation

Kevelson, M. J. C. (2019). *The measure matters: Examining achievement gaps on cognitively demanding reading and mathematics assessments* (Research Report No. RR-19-43). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12278

Action Editor: Elizabeth Stone

Reviewers: Katherine Castellano and Samuel Rikoon