

Test Score Gaps in Higher Order Thinking Skills: Exploring Instructional Practices to Improve the Skills and Narrow the Gaps

Hajime Mitani

Rowan University

Today's economy demands higher order thinking skills (HOTS), and the public education system has a critical role in supporting students' acquisition of HOTS. Yet, numerous studies documented inequity in access to higher quality instruction that promotes HOTS, which could result in wide test score gaps in HOTS. In this study, I examined test score gaps in HOTS and explored instructional practices associated with HOTS, particularly among low-performing students, using large-scale international assessment data from the 2015 Trends in Mathematics and Science Study. I found wide test score gaps in HOTS in mathematics between the lowest and highest socioeconomic status students and between White students and students of color. Instructional practices such as the same ability group work, asking students to work on problems with teacher guidance, and working on problems with no immediately obvious method of solution were found positively associated with the test scores.

Keywords: *test score gaps, higher order thinking skills, engaging instructional practices*

TODAY'S economy demands higher order thinking skills (HOTS) for innovation and development, especially in the science, technology, engineering, and mathematics (STEM) areas. A recent report by the World Economic Forum (2015) showed that jobs requiring nonroutine interpersonal and analytical skills have increased gradually, while jobs requiring routine manual and cognitive skills declined steadily between 1960 and 2010. The Pew Research Center (2016) also reported that jobs that require higher levels of analytical skills have increased by 77% since 1980. Many companies, organizations, and media now call for 21st-century skills, which generally include critical thinking skills, problem-solving skills, information and communication technology literacy, and leadership skills (Partnership for 21st Century Learning, n.d.; World Economic Forum, 2015),¹ and economic and social returns to these skills are likely to increase in near future. HOTS can be viewed as a cognitive dimension of 21st-century skills and relates to critical thinking skills and problem-solving skills. Not possessing HOTS may, therefore, put one in a disadvantaged position in the future labor market.

The public education system has a critical role in supporting students' acquisition of HOTS because these skills can be enhanced through student-centered instruction and are transferrable after they exit PreK–12 education systems (Guskey, 2007; Halpern, 1998; Hmelo & Ferrari, 1997). In fact, 60% of Americans think that public schools should have a vast responsibility for ensuring that future workers have the right skills to succeed in today's and future economy (Pew Research Center, 2016).

Unfortunately, not all students receive adequate education to acquire HOTS. It is well documented that teacher quality is inequitably distributed across schools (e.g., Clotfelter et al., 2005; Clotfelter et al., 2007; Goldhaber et al., 2015; Hanushek et al., 2004; Lankford et al., 2002). This educator sorting contributes to inequity in instructional quality because low socioeconomic status (SES) schools and schools serving large proportions of students of color tend to have inexperienced teachers who are new to the profession, often struggle with foundational teaching skills, and lack the skills to implement student-centered learning (Goldhaber et al., 2015; Maas et al., 2018; Mehta, 2014; Noguera et al., 2015).

Academic tracking and instructional time may be other factors that contribute to unequal access to HOTS. Research shows that low-SES students and students of color are more likely to be placed in basic classes, which focus on mastery of low-level skills, and receive less instructional time (Burris & Welner, 2005; Gamoran, 1987; Mickelson, 2001; Oakes, 2005; Rogers et al., 2014).

Furthermore, students' home environments could play a significant role in their readiness to learn from higher order thinking (HOT) activities. Research shows that students from low-SES families have fewer opportunities to HOT activities (e.g., Kalil, 2015; Kalil et al., 2016; Lareau, 2003; Phillips, 2011). It also documents that conversations between students of color and their parents tend to be directive and authoritarian, which provides the students less time to engage in HOT activities (Lareau, 2003).

Students studying at these schools and from such home environments could be exposed differentially to instructional



practices that put the students in the center of learning and demand high cognitive processes. While prior research generally suggests that they benefit from such instructional practices (e.g., Boaler, 2002; L. M. Martin & Halpern, 2011; Zohar & Dori, 2003), it is theoretically possible they learn differently by student subgroup and type of instructional practices. As a result of these differences in school and home environments, test score gaps in HOTS may exist among student subgroups. Likely, students respond differently to instructional practices that engage students in HOT activities.

Surprisingly, most of the prior work focused on subject-level test score gaps and have not explored the gaps in HOTS explicitly or the factors that relate to improvements in HOTS among low-performing student subgroups (e.g., Curran & Kellogg, 2016; Fryer & Levitt, 2004; Hanushek & Rivkin, 2006; Reardon, 2011; Reardon & Galindo, 2009). To advance our understanding of the test score gaps and instructional practices that are associated with HOTS, this study examines the gaps among student subgroups in eighth-grade mathematics and explores instructional practices that may be linked to HOTS using large-scale international assessment data from the 2015 Trends in International Mathematics and Science Study (TIMSS). It focuses on mathematics because the number of STEM occupations has been growing much faster than non-STEM occupations (Fayer et al., 2017), and mathematics skills are essential to succeed in STEM fields (National Council of Teachers of Mathematics, 2018). If the test score gaps in HOTS in mathematics persist, it is likely to increase inequity and inequality in later adult outcomes.

More formally, this study explores three research questions. First, it examines test score gaps in HOTS by SES status and race/ethnicity. I hypothesize that White students and high-SES students outperform their counterparts. Second, it investigates instructional practices that may be related to higher test scores in HOTS. I hypothesize that student-centered instructional practices are associated with higher test scores in HOTS. Third, it examines whether the relationships between these instructional practices and the test scores vary by students' SES status and racial/ethnic backgrounds. I hypothesize that such variation exists. The third question aims to identify instructional practices for each student subgroup, particularly the lowest SES students and students of color, that relate to higher test scores in HOTS.

For the third research question, since HOT activities demand time and necessitate students' prior and current knowledge, instructional time, and content coverage could moderate the relationships. This is a plausible hypothesis based on Carroll's (1963) school learning model and prior research (de Jong & Lazonder, 2014; Etkina et al., 2008; Luyten, 2017; Masek & Yamin, 2011; Polikoff & Porter, 2014; Zimmerman, 2007). I examine whether a moderation

effect exists by instructional time and/or content coverage. Due to the space limitation, I reported and discussed the results in the online supplement.

It is important to note that I use the term, *test score gaps*, throughout the article instead of the commonly used term, *achievement gaps* because the latter term has negative connotations that treat White, affluent student achievement as a norm (e.g., Love, 2004). The use of this language may affect the priorities of educators and lead us to develop short-term solutions that do not address the root causes of the problem (Ladson-Billings, 2006; Quinn et al., 2019). It is also important to note that many other factors contribute to the test score gaps. Disentangling their effects is methodologically challenging and beyond the scope of this study. Instead, the current study analyzes the gaps and the associations between instructional practices and test scores more descriptively and does not intend to make a causal inference. The results point to the direction for further investigation in which researchers may use a more rigorous research design to estimate a causal effect.

This article proceeds as follows. The first section describes HOTS and discusses instructional practices that could be theoretically related to HOTS and some sources of test score gaps in HOTS. The next section reviews prior studies on test score gaps. The following section describes the TIMSS assessment data and explains the methods used. After the method section, it reports findings and concludes with discussions and implications.

Higher Order Thinking Skills

Many people use the term, HOTS, to describe some form of complex thinking that demands high cognitive processes. It was a term developed to merge two different perspectives about critical thinking from the field of philosophy, which viewed it as evaluation or judgment, and from the field of psychology, which viewed it as problem solving (Lewis & Smith, 1993). HOT is defined as thinking that occurs "when a person takes new information and information stored in memory and interrelates and/or rearranges and extends this information to achieve a purpose or find possible answers in perplexing situations" (Lewis & Smith, 1993, p. 136). As the historical development of the term indicates, HOT includes evaluation or judgment, problem solving, as well as creative thinking, and decision making (Lewis & Smith, 1993). Since HOTS results from a merger of the two different perspectives on critical thinking, it also includes broadly defined critical thinking as its component. Reasoning or productive thinking is part of problem solving, as it is used to integrate past experiences that have not been associated to find a solution to a novel challenge (Lewis & Smith, 1993).

Other researchers, more recently, define HOTS differently but their definitions generally capture similar components. For example, Schraw et al. (2011) view HOTS as thinking

Bloom's Taxonomy

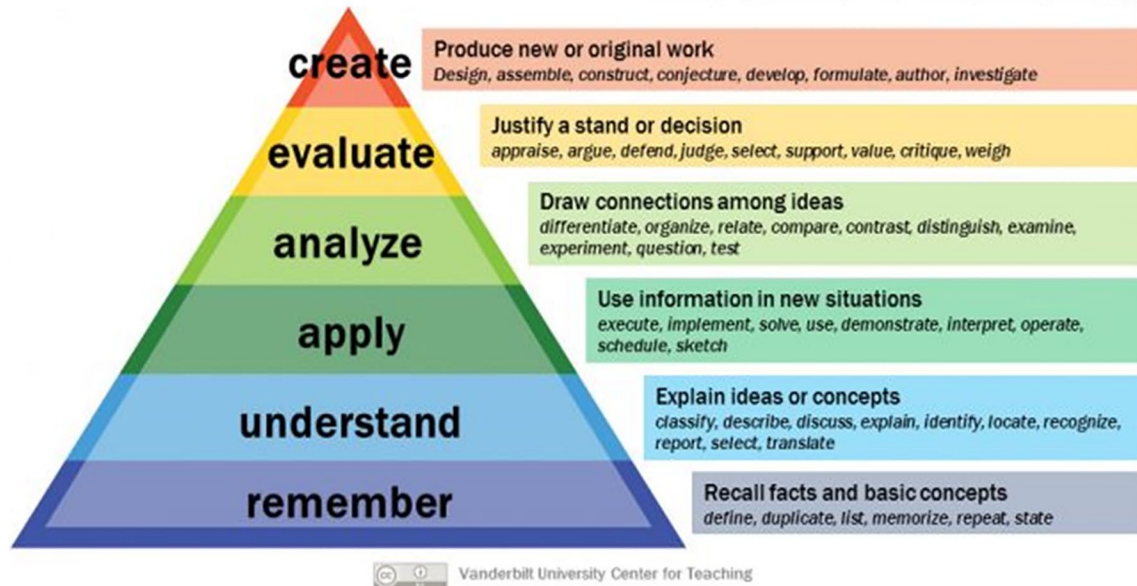


FIGURE 1 *Bloom's taxonomy.*
Source. Vanderbilt University Center for Teaching.

skills composed of reasoning, argumentation, problem solving and critical thinking, and metacognition. Brookhart (2010) defines HOTS in terms of transfer (i.e., making sense of and being able to use what one has learned), critical thinking, and problem solving. Richland and Simms (2015) argue that the underlying mechanism across these components is relational or analogical reasoning, which is “the process of representing information and objects in the world as systems of relationships, such that these systems of relationships can be compared, contrasted, and combined in novel ways depending on contextual goals” (Richland & Simms, 2015, p. 177).

Bloom's original and revised taxonomy provides additional inputs about what HOTS means. It classifies learning objectives into major categories in the cognitive process dimension (Krathwohl, 2002). The revised framework, as shown in Figure 1, includes the following six categories: *remember*, *understand*, *apply*, *analyze*, *evaluate*, and *create* (see Bloom et al., 1956, for the original taxonomy, and Anderson et al., 2001, for the revised one). The framework also includes the knowledge dimension, which is foundational to perform the six cognitive processes. These process categories are ordered in terms of their complexity with *remember* being the least complex and *create* being the most complex; yet the framework is not strictly hierarchical (Ennis, 1993). The categories are allowed to overlap and theoretically interdependent (Ennis, 1993; Krathwohl, 2002). The upper three levels are closely aligned with several definitions offered earlier and generally considered as

HOTS (Richland & Simms, 2015; Tankersley, 2005, as cited in Hitchcock, 2020; Zohar & Dori, 2003). The definition and example of each of these cognitive processes are provided in Appendix Table 1, which is part of Table 5.1 in Anderson et al. (2001).

HOTS Specific to Mathematics

In mathematics, HOTS is implicitly embedded in the Common Core State Standards for Mathematics (CCSSM).² The CCSSM includes content and practice standards for mathematics, and the latter standards describe eight cognitive processes (Common Core State Standards Initiative, n.d.). The practice standards include standards in the following eight domains: (1) *make sense of problems and persevere in solving them*; (2) *reason abstractly and quantitatively*; (3) *construct viable arguments and critique the reasoning of others*; (4) *model with mathematics*; (5) *use appropriate tools strategically*; (6) *attend to precision*; (7) *look for and make use of structure*; and (8) *look for and express regularity in repeated reasoning*. All of them are closely related to HOTS. Richland and Begolli (2016) offer a detailed analysis of these processes concerning analogical reasoning, which they argue is the underlying mechanism of HOTS.

Instructional Practices Related to Students' HOTS

The literature on cognitive psychology and teaching and learning provides scientific evidence on instructional

practices that are positively associated with students' HOTS. One approach is a classroom and small-group discussion and dialogue. Studies found that dialogue and discussions can promote student learning and help students develop HOTS (e.g., Abrami et al., 2015; King, 1992; Miri et al., 2007; Murphy et al., 2016; Slavin, 2011; Snyder & Snyder, 2008). In this approach, teachers encourage students to discuss their thoughts with their classmates and explain their answers and reasons. This provides the students with opportunities to think deeply about concepts; integrate them with their prior knowledge; evaluate and assess their thoughts, assumptions, and understandings; and make a decision about what to believe or do (Bailin et al., 1999; Bonney & Sternberg, 2011; King, 2002). This higher order metacognitive process helps them construct new knowledge. Teachers may facilitate the discussion through sets of questioning strategies such as the *Ask to Think Tel-Why* model, the *Cognitive Tools and Intellectual Roles* approach, and the *Guided Reciprocal Peer Questioning* approach (Gillies, 2011; King, 2002).

Teachers may use mixed or same ability grouping for small group discussions. The literature provides mixed evidence about the effectiveness of ability grouping. A recent meta-analysis found that students learn more when they are grouped by ability within the classroom (Steenbergen-Hu et al., 2016). The mixed-ability grouping may enhance students' HOTS as well if it provides students opportunities to experience HOT activities. Yet, if group work is designed such that a single group product is required and/or no group reward/incentive is provided, less able students may not learn very much because more able students do most of the work (Slavin, 2011).

Another approach teachers may use is inquiry-based learning. In this approach, teachers provide their students enough time and adequate scaffolding/guidance to explore a topic or work on a problem within parameters set by the teachers (Marshall & Horton, 2011). During the self-directed investigation, the students may draw on their prior knowledge and understanding (Marshall & Horton, 2011). Inquiry-based learning is found correlated with high-level cognitive thinking (Lazonder & Harmsen, 2016; Marshall & Horton, 2011). The effectiveness of inquiry-based learning depends on whether the teachers provide appropriate scaffolding/guidance (Lazonder & Harmsen, 2016).

Problem-based learning (PBL) is also found positively associated with HOTS if it provides well-designed problems and adequate scaffolding (e.g., Hung et al., 2008; Weiss, 2003). In PBL, students work on a complex, real-world problem where no immediately obvious method of solution is available. Such a problem requires knowledge slightly beyond the students' current knowledge and engages them in collaborative inquiry in small groups (Lu et al., 2014; Weiss, 2003). The students initiate their prior knowledge and apply it to the problem, discuss the nature of the problem with their peers

and identify knowledge gaps, evaluate and synthesize proposed ideas, solve controversies and make decisions based on the discussion, and refine their current knowledge and construct new understanding (Lu et al., 2014; Weiss, 2003).

These specific approaches use multiple instructional practices that do not have specific labels but are key to engaging students in active learning. Some of them include relating the lesson to students' daily lives, asking them to explain their answers, encouraging them to express their ideas, and linking new content to their prior knowledge. Research shows that these practices are positively related to HOTS. For example, Miri et al. (2007) found that students improved their HOTS when they dealt with real-world problems, engaged in open-ended discussions, and experienced inquiry-oriented experiments. Slavin (2011) argues that one of the most effective ways that facilitate the elaboration of the content is to have students explain the content to someone else.

These and other instructional practices not described above necessitate prior knowledge, and students with such knowledge would benefit more from HOT activities (Richland & Simms, 2015). This does not mean that students with less prior knowledge do not benefit from such activities. Research shows evidence that students develop HOTS, whether their initial academic skill levels are low, average, or high (e.g., Crosnoe et al., 2010; Zohar & Dori, 2003), suggesting that a mastery of basic skills may not be always necessary to acquire HOTS.

Instructional Time and Content Coverage

Carroll's (1963) school learning model posits that the degree of student learning is a function of time needed for learning and time spent on learning, the former of which is a function of aptitude, ability to understand instruction, and quality of instruction. The latter is a function of time allowed for learning and perseverance. The model suggests that, while students' abilities, attitudes, and dispositions play a role in their learning, students may generally learn more if instructional time increases. Given that HOT activities require a good amount of time, this interactive effect is plausible. Prior research and reviews suggest this possibility (e.g., de Jong & Lazonder, 2014; Etkina et al., 2008; Masek & Yamin, 2011; Zimmerman, 2007).

Another important factor that arises from Carroll's model is an opportunity to learn. Although time allowed for learning is labeled as *opportunity to learn* in his model, the concept now involves more than a simple time dimension (Floden, 2002). A commonly used definition focuses on content coverage or the extent of student exposure to assessed topics (e.g., Scheerens, 2017; Schmidt et al., 2015; Schmidt et al., 2019).³ Limited exposure to content results in less prior and current knowledge, which may affect the effectiveness of the instructional practices (e.g., Richland & Simms, 2015).

Similar to instructional time, this suggests a possible interactive relationship between the instructional practices and content coverage (e.g., Luyten, 2017; Polikoff & Porter, 2014).

(Some) Sources of Test Score Gaps in HOTS

Test score gaps in HOTS could result from systematic differences in students' school and home environments, as underscored in the ecological framework (Bronfenbrenner, 1979, 1986). An in-depth discussion of each possible source is beyond the scope of the current study. Instead, I focus on some of the important factors that are likely to be associated with the gaps in HOTS.

The first factor is teacher experience and sorting. Some studies found that less experienced teachers tend to ask students more factual and lower order thinking questions; other studies reported that new teachers often struggle with classroom management (e.g., Castro et al., 2010; Dias-Lacy & Guirguis, 2017; He & Cooper, 2011). Since students of color and low-SES students are more likely to be assigned to less experienced teachers (e.g., Clotfelter et al., 2005; Clotfelter et al., 2007; Goldhaber et al., 2015; Hanushek et al., 2004; Lankford et al., 2002), their opportunities for HOT activities may be limited.

Another possible source of the gap is teacher expectation. HOT demands high-level cognitive processes. Some teachers think that it is less appropriate for low-performing students because they perceive that learning occurs hierarchically and HOT occurs after mastering prerequisite skills (Zohar et al., 2001). They tend to believe that HOT activities are more effective with high-SES, high-achieving students (Warburton & Torff, 2005). If teachers collectively hold low expectations for HOTS among students of color and low-SES students, it forms a poor academic culture at the school, and the students may suffer from instruction that rarely requires HOTS.

Academic tracking deprives students of color and low-SES students of opportunities for HOT activities. Research shows that they are more likely to be placed into basic classes and programs, and their schools provide them limited access to advanced classes (Burris & Welner, 2005; Gamoran, 1987; Ladson-Billings, 1997; Mickelson, 2001; Oakes, 2005; Patrick et al., 2020). Teachers in these classes and programs tend to use traditional instructional practices and provide them fewer opportunities for HOT activities (Darling-Hammond, 2001; Desimone & Long, 2010; Ladson-Billings, 1997; Noguera et al., 2015). This suggests that teachers' use of the instructional practices discussed earlier reflect academic tracking.

Schools serving large proportions of students of color and low-SES/income students generally struggle with implementing instructional practices for HOT (Maas et al., 2018; Noguera et al., 2015). These schools tend to be low-performing schools and have to shift resources to prepare

students for standardized testing, which does not necessarily assess students' HOTS (Au, 2007; He & Cooper, 2011; Noguera et al., 2015). Generally, teachers at these schools tend to be less effective at providing rigorous and engaging instruction, particularly for students with weaker academic and social-emotional skills (Maas et al., 2018).

Students' home environments also play a role. Prior research shows that children from high-SES/income families are exposed to a variety of cognitive activities daily at home that may influence the development of HOTS. For example, their parents read books to their children when they are young, have language-rich conversations and ask HOT questions, and oversee homework completion (e.g., Altintas, 2016; Hart & Risley, 1995; Kalil, 2015; Wilder, 2014). On the other hand, low-SES/income parents tend to provide their children fewer cognitively stimulating activities at home (e.g., Kalil, 2015; Kalil et al., 2016; Lareau, 2003; Phillips, 2011). These differences in home environments could lead to test score gaps in HOTS. A recent study shows that family income affects the onset and the development trajectories of HOT talk among children between 14 months and 58 months (Frausel et al., 2020). It found that children from higher income families start using HOT talk earlier than those from lower income families. It also found that family income, parent education, and parent IQ are positively correlated with children's HOT outcomes in grade school.

These systematic differences in students' school and home environments could be linked to possible differences in students' readiness for HOT activities. One student subgroup may learn more from a given instructional practice than other subgroups. Similarly, within subgroups, some instructional practices may be more strongly related to students' HOTS than other practices.

Studies on Test Score Gaps

Myriad researchers and organizations have reported subject-level test score gaps among student subgroups since the 1966 Coleman Report. Although the magnitude of the gaps varies from study to study, they generally found low-SES students, students of color, and English language learner (ELL) students underperform their counterparts in standardized tests in mathematics, reading, and science at all grade levels (e.g., Clotfelter et al., 2009; Duncan & Magnuson, 2011; Fryer & Levitt, 2004; Hemphill & Vanneman, 2011; Jencks & Phillips, 1998; Reardon, 2011; Reardon & Galindo, 2009; Rumberger & Tran, 2010). For example, Clotfelter et al. (2009) found that the raw test score gap in Grade 3 between Black and White students was 0.78 SD in mathematics and 0.71 SD in reading. The gaps remained sizable even after regression-based adjustments. Reardon (2011) and Duncan and Magnuson (2011) reported that the test score gaps by income level and SES are more than 1 SD,

which is roughly equivalent to 3 to 6 years of learning. The gap was more than double the size of the Black–White test score gap. The gaps between ELL students and non-ELL students are smaller and less than 1 SD (Hemphill & Vanneman, 2011; Reardon & Galindo, 2009; Rumberger & Tran, 2010).

These gaps already exist when children enter kindergarten and even among toddlers and preschoolers in terms of vocabulary and language development. For instance, toddlers raised by lower SES families are 6 months behind toddlers from higher SES families in language proficiency during the first 24 months (Fernald et al., 2013). Low-income children are exposed to fewer vocabulary words than high-income families, contributing to the language gap (Hart & Risley, 1995). When these children start kindergarten, they score 1.3 SD lower than those from low-need/higher SES families in entry mathematics skill assessments (Duncan & Magnuson, 2011). In science, a somewhat smaller but still sizable gap is observed between kindergarteners from lower and higher income families (Curran, 2017). By race and ethnicity, the test score gap is 0.82 SD between Black and White students and 0.94 SD between Hispanic and White students (Curran & Kellogg, 2016).

Compared with the volume of research on the subject-level test score gaps, test score gaps in HOTS have received scant attention in the literature. Such information is essential for practitioners to design better academic programs. The current study fills this gap in the literature.

Data

In this study, I used international large-scale assessment data from the TIMSS 2015, from which I extracted U.S. assessment data in eighth-grade mathematics.⁴ TIMSS is a repeated cross-sectional international large-scale assessment study conducted by the International Association for the Evaluation of Educational Achievement every 4 years since 1995. The United States has participated in the study for all cycles since 1995.⁵

TIMSS uses a stratified two-stage cluster sampling design with schools being the first sampling unit and intact classrooms being the second sampling unit (M. O. Martin & Mullis, 2012). In each administration, it samples students at Grade 4 and Grade 8 separately and assesses their cognitive skills in mathematics and science at the subject level and two domain levels (i.e., content and cognitive domains). The cognitive domain in both grades and subjects includes the following three cognitive domains: *knowing*, *applying*, and *reasoning* (Mullis et al., 2009).⁶

In addition to the assessment, TIMSS collects contextual factors through surveys of students, their teachers, and their schools. All of these data can be merged into a single data file. For this study, after combining all TIMSS data files, I merged them with the 2014–15 school-level Common Core of Data (CCD) from the U.S. Department of Education to

incorporate school information on the percentage of students eligible for the federal free/reduced lunch program, which was used to construct a socioeconomic status variable.⁷

The sample consisted of 9,630 eighth-grade students in 500 classrooms taught by 390 math teachers at 230 public schools.⁸ Students in private schools were excluded from this study because the focus was on whether the U.S. public education system is equitable in terms of teaching students HOTS and preparing them to be successful in the future economy.

TIMSS Assessments

The TIMSS assessment uses a matrix-sampling approach, in which the entire pool of assessment items at each grade level is divided into 14 booklets for each subject, and each student completes one assessment booklet per subject (Mullis et al., 2009).⁹ Assessment items include multiple-choice and constructed-response questions and at least one half of the total number of points come from multiple-choice questions. The assessment takes 90 minutes, and additional 30 minutes are spent on the questionnaire at the eighth grade (Mullis et al., 2009).

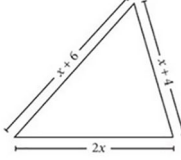
Because students' abilities are measured based on a small set of assessment items, there is a substantial amount of measurement error. To address this problem and obtain unbiased parameter estimates, the TIMSS uses plausible values, which represent the likely distribution of a student's academic ability (e.g., von Davier et al., 2009). The plausible values are randomly drawn from the posterior distributions, five times for each student. I used all of the five plausible values to estimate parameters through multiple imputation techniques with complex survey weights (see Allison, 2002; Little & Rubin, 2002). Standard errors were estimated through the Jackknife repeated replication variance estimation method.

Cognitive Domain in TIMSS

The cognitive domain consists of three domains: *knowing*, *applying*, and *reasoning* (Mullis & Martin, 2013). The *knowing* domain includes six categories: *recall*, *recognize*, *classify/order*, *compute*, *retrieve*, and *measure*. These processes require relatively simple cognitive processes and do not demand HOT. The *applying* domain is composed of three categories: *determine*, *represent/model*, and *implement*. In this domain, students apply mathematical facts, concepts, and procedures they already understand to real-life situations or purely mathematical questions they are familiar with and solve problems (Mullis & Martin, 2013). Problem solving is central in this domain and part of HOTS; yet, the scope of this domain is limited to the application of knowledge to *familiar* situations and problems, rather than *complex*, *novel* situations. In this sense, this cognitive process

(a)

Content Domain: Algebra
Cognitive Domain: Applying
Description: Constructs and uses the solution of a linear equation to solve a word problem



The sum of the lengths of the sides of this triangle is 30 cm.

A. Write an equation that would enable you to find the value of x .


Equation: $4x + 10 = 30$

B. What is the length of the LONGEST side of the triangle in centimeters?

Answer: 11 cm

(b)

Content Domain: Geometry
Cognitive Domain: Reasoning
Description: Uses the Pythagorean theorem in finding the perimeter of a trapezoid



$ABCD$ is a trapezoid with $AB = 10$ cm and $CD = 16$ cm. $AD = BC$. The distance between the parallel lines, AB and CD , is 4 cm. What is its perimeter?

☒ A 36 cm
☐ B 34 cm
☐ C 32 cm
☐ D 30 cm

FIGURE 2 Sample assessment item in the (a) applying domain and (b) reasoning domain.

Source. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

does not fully capture HOTS. I used test scores in the *knowing* and *applying* domains as references.

The last cognitive domain, *reasoning*, includes the following six categories: *analyze*, *integrate/synthesize*, *evaluate*, *draw conclusions*, *generalize*, and *justify*. These cognitive processes match with HOTS that I described earlier and the CCSSM very well. I used test scores in the *reasoning* domain as HOTS in this study.¹⁰ Figure 2 shows a sample of eighth grade mathematics questions in the *applying* and *reasoning* domains. Appendix Table 2 provides more detailed information about each cognitive process taken from the TIMSS 2015 Mathematics Framework (see Grønmo et al., 2013).

Variables

Race/Ethnicity. I used a race/ethnicity variable to create indicator variables for students' race/ethnicity. It includes the following seven race/ethnicity categories: *White (not Hispanic)*, *Black (not Hispanic)*, *Hispanic*, *Asian*, *Native American*, *Pacific Islander*, and *two or more races*. I created indicator variables for the first four race/ethnicity categories and collapsed the rest into the *other race/ethnicity* category.

Socioeconomic Status. The TIMSS 2015 data files do not include a variable for students' SES, so I created it using a set of variables available in the data files. Researchers suggest that SES components should include family income, parental educational attainment, and parental occupational status as well as neighborhood and school SES (Brunner, 2014; Cowan et al., 2012). Unfortunately, the TIMSS data files are limited regarding direct measures of some of these SES components. I constructed an SES variable using variables that can be used as proxies for some aspects of these SES components.

First, for family income, I used survey items related to possessions of such items as books, computers, internet connection, and own room. I also used survey items related to activities outside of school such as playing on a sports team, playing a musical instrument, and belonging to a club, because students often cannot engage in these activities without adequate financial capital. I did not use parental educational attainment, since more than 20% of the data were missing.¹¹ For school and neighborhood SES, I used a school-level variable on the percentage of students eligible for the federal free/reduced lunch program from the CCD data file.

To construct a single SES variable, I used principal component analysis based on polychoric correlations. I created a single, standardized SES variable and divided it into quintiles.¹² Appendix Table 3 provides a complete list of survey items and data used to create this variable.

Instructional Practices Related to HOTS. TIMSS data include sets of questions implicitly and explicitly related to the instructional practices described earlier. One set of questions asked teachers how often they relate the lesson to students' daily lives, ask students to explain their answers, ask students to complete challenging exercises that require them to go beyond the instruction, encourage classroom discussions among students, link new content to students' prior knowledge, ask students to decide their problem-solving procedures, and encourage students to express their ideas in class. These instructional practices underlie the specific instructional approaches described earlier and the CCSSM. I reverse-coded teachers' responses with 1 being *never* and 4 being *every or almost every lesson*, took the mean across the seven questions, and labeled it as *engaging instructional practices*.¹³ Appendix Table 4 reports the means and standard deviations of these seven survey items.

Another set of questions asked teachers how often they ask students to work on problems (individually or with peers) with their guidance, work on problems for which there is no immediately obvious method of solution, work in mixed ability groups, and work in same ability groups. The first two practices relate to inquiry-based learning with scaffolding/guidance and PBL¹⁴; and the last two practices are concerned with small group learning. Similar to the first set

of questions, for each practice, I reverse-coded teachers' responses with 1 being *never* and 4 being *every or almost every lesson*. Note that these variables only measure the *frequencies* of these practices, not the quality.

For instructional time, I used a variable that measures the number of minutes spent on teaching mathematics to the students in the TIMSS class per week. I transformed minutes into hours. For content coverage, I used teachers' responses to sets of questions about content coverage in all four content areas in the content domain (i.e., *number*, *algebra*, *geometry*, and *data and chance*). Each content area has multiple topics, and teachers were asked to indicate whether a topic in each content area was mostly taught before this year (i.e., 2014–2015), was mostly taught this year, or was not yet taught or just introduced. There are 20 topics across the four content areas. I created an indicator variable for each topic that takes a value of one if the topic was taught before this year or mostly taught this year and a value of zero otherwise. Then, I took the average of the 20 indicator variables and expressed it as a percentage of content in the TIMSS assessment taught by this year.¹⁵

Student and Teacher Characteristics. In the analysis that follows, I also used sets of basic student and teacher characteristics as controls. Student characteristics include age, sex, the number of days absent from school, how often English is spoken at home, ever repeated a grade in elementary school, ever repeated a grade in middle or junior high school, special accommodation provided during the mathematics assessment, and students' confidence in mathematics. Student confidence relates to aptitude, ability to understand instruction, and perseverance in Carroll's (1963) model. The confidence variable was constructed by the TIMSS project staff based on the Rasch partial credit model.¹⁶ To remove the influence of academic programs and support not provided by the school, I used an indicator variable for extra mathematics lessons and tutoring.

Teacher characteristics include age, sex, highest educational attainment, years of teaching experience, college major, the number of hours spent on professional development in the past 2 years, the number of content areas that professional development covered, and job satisfaction. The job satisfaction variable was created by the TIMSS project staff based on the Rasch partial credit model.

As explained in the method section, I did not use school-level variables. Instead, I used school fixed effects to control for both observable and unobservable school-level factors. Table 1 reports descriptive statistics on students and teachers in the analytic sample.

Method

I first estimated raw test score gaps in the *reasoning* domain by race/ethnicity and SES separately through ordinary least squared regression techniques and then reestimated

the gaps including both characteristics in a single linear regression model.^{17, 18} After that, I reestimated the gaps, controlling for student characteristics. The model takes the following form:

$$A_i = \beta_0 + \beta_1 SESQ1_i + \beta_2 SESQ2_i + \beta_3 SESQ3_i + \beta_4 SESQ4_i + \beta_5 Black_i + \beta_6 Hispanic_i + \beta_7 Asian_i + \beta_8 OtherRace_i + \gamma St_i + \varepsilon_i \quad (1)$$

where A_i is a standardized test score of the i th student, each of the SES quintile and race/ethnicity variables is an indicator variable, St_i is a vector of student characteristics of the i th student, and ε_i is an unobserved random error term. A reference group for SES status was the fifth quintile, and a reference group for race/ethnicity was White students.

Next, to estimate the relationships between the instructional practices and the test scores in the *reasoning* domain, I estimated a series of regression models that include instructional practices, instructional time, content coverage, student controls, teacher controls, and school fixed effects. I included school fixed effects to remove the effect of observable and unobservable between-school factors that affect both the instructional practices and the test scores. In this sense, I utilized variation in the instructional practices within schools to estimate the associations.¹⁹ The main model takes the following form:

$$A_{ij} = \beta_0 + \beta_1 SESQ1_{ij} + \beta_2 SESQ2_{ij} + \beta_3 SESQ3_{ij} + \beta_4 SESQ4_{ij} + \beta_5 Black_{ij} + \beta_6 Hispanic_{ij} + \beta_7 Asian_{ij} + \beta_8 OtherRace_{ij} + \gamma St_{ij} + \delta IP_{ij} + \theta_j + \varepsilon_{ij} \quad (2)$$

where IP is a vector of instructional practice variables, instructional time, and content coverage of the i th student in the j th school, and θ_j is school fixed effects.^{20, 21}

Results

Test Score Gaps in Reasoning in Eighth-Grade Mathematics

Figure 3 plots the coefficient estimates and their 95% confidence intervals on the SES and race/ethnicity variables from the four sequential linear regression models described in the method section. Figure 3a plots estimates on the SES quintile variables; Figure 3b plots estimates on the race/ethnicity variables; and the last two figures (c) and (d) plot estimates on the SES quintile variables and the race/ethnicity variables. The estimates in Figure 3d come from Equation 1.

These figures show sizable test score gaps in all cognitive domains. For example, the raw gaps between the lowest and highest SES students were all close to 1 SD. The gaps still remained sizeable and were close to one half of one SD even after regression adjustments. The raw gaps between Black and White students ranged from 0.80 SD to 0.91 SD and remained in the range of 0.61 SD and 0.70 SD after regression adjustments. The raw gaps between Hispanic and White

TABLE 1
Descriptive Statistics

	<i>M</i>	<i>SD</i>
Student characteristics		
Age	14.20	0.45
Girl	0.51	
SES, first quintile	0.17	
SES, second quintile	0.19	
SES, third quintile	0.20	
SES, fourth quintile	0.22	
SES, fifth quintile	0.22	
White, not Hispanic	0.51	
Black, not Hispanic	0.11	
Hispanic	0.26	
Asian	0.05	
Other race/ethnicity category	0.07	
School absence (1 = <i>once a week or more</i> ; 4 = <i>never or almost never</i>)	3.48	0.80
English spoken at home (1 = <i>never</i> ; 4 = <i>always</i>)	3.66	0.67
Ever repeated a grade in elementary school	0.06	
Ever repeated a grade in middle school	0.01	
Special accommodation provided	0.07	
Confident in mathematics (standardized)	0.03	1.00
Attended extra mathematics lessons or tutoring not provided by the school	0.24	0.43
Teacher characteristics		
Age (1 = below 25; 2 = 25–29; 3 = 30–39; 4 = 40–49; 5 = 50–59; 6 = 60 years or older)	3.64	1.27
Female	0.69	
Educational attainment (1 = HS dropout; 6 = doctorate)	4.61	0.51
Years of teaching experience	14.04	9.92
Majored in mathematics	0.50	
Majored in biology	0.03	
Majored in physics	0.04	
Majored in chemistry	0.02	
Majored in earth science	0.02	
Majored in mathematics education	0.58	
Majored in science education	0.09	
Majored in education	0.36	
Majored in other area	0.39	
Professional development hours (1 = none; 5 = more than 35 hours)	3.69	1.16
Number of content areas PD covered (0–7)	4.98	2.01
Job satisfaction (standardized)	0.00	1.01
Instructional practice		
Engaging instructional practices–Averaged	3.36	0.51
Small group work–Same ability group*	2.34	0.80
Small group work–Mixed ability group*	2.89	0.86
Work on problems with teacher guidance*	3.57	0.65
Work on problems with no immediately obvious method of solution*	2.33	0.80
Total hours spent on mathematics instruction per week	4.36	1.50
Percentage of content in the TIMSS assessment taught by this year	90.56	11.57

Note. SES = socioeconomic status; TIMSS = Trends in International Mathematics and Science Study. An asterisk indicates the scale with 1 being *never* and 4 being *every or almost every lesson*. These statistics were estimated based on the analytic sample used for the subsequent analyses. An unconditional approach was used (West et al., 2008). Student and math teacher weights (linear transformation of total student weights) were used to estimate means and standard deviations. Although not reported, standard errors were estimated by the Jackknife repeated replicate sampling variance estimation method. By survey design, statistics on teachers do not represent a teacher population.

Source. U.S. Department of Education, National Center for Education Statistics, TIMSS 2015.

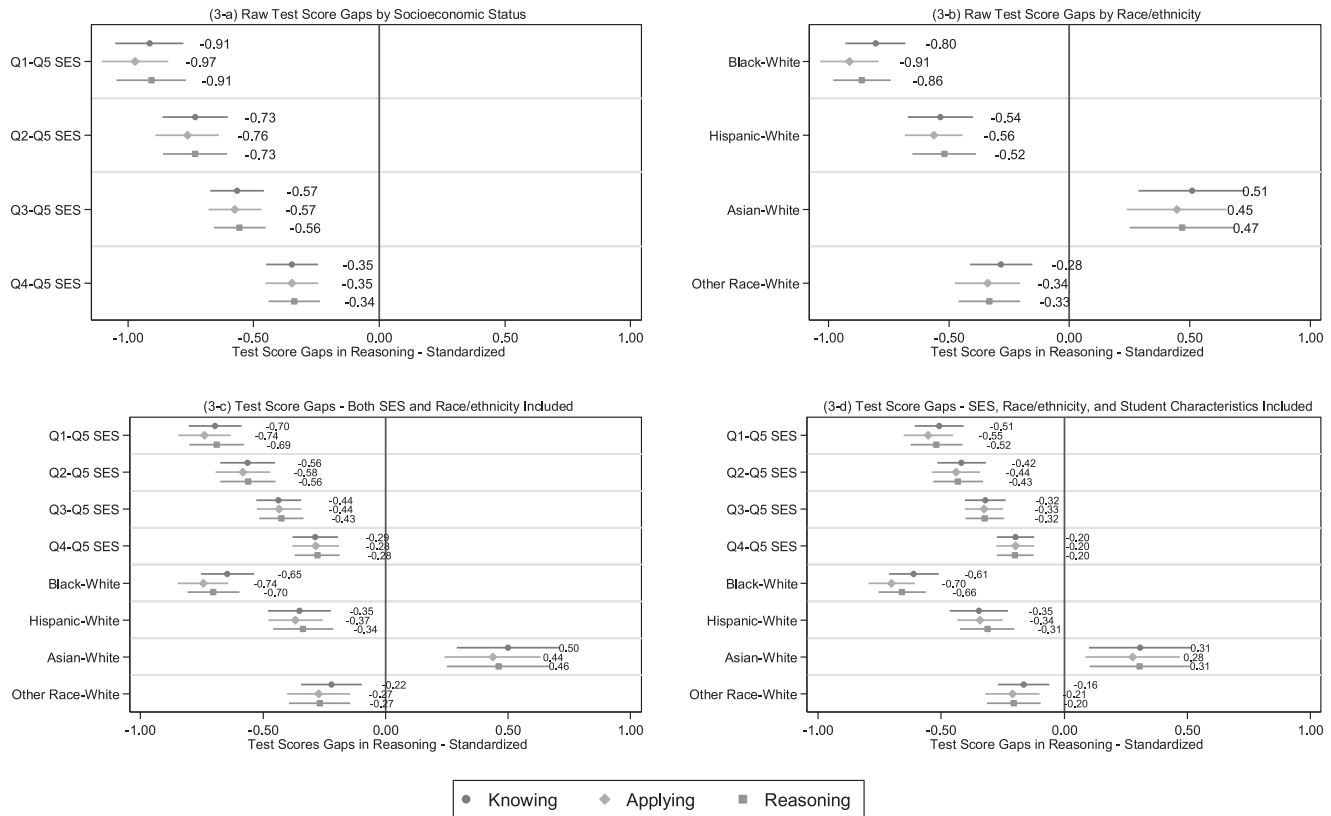


FIGURE 3 Test score gaps in the reasoning domain.

Note. Q1 = first quintile; Q2 = second quintile; Q3 = third quintile; Q4 = fourth quintile; Q5 = fifth quintile. Q1 – Q5 SES shows the coefficient on the indicator variable for the first SES quintile, which indicates the mean difference in the test score in each cognitive domain between the students in the first quintile and fifth quintile. Black–White shows the coefficient on the indicator variable for Black students, which indicates the mean difference between Black and White students.

Source. U.S. Department of Education, National Center for Education Statistics, Trends in International Mathematics and Science Study 2015.

students were the smallest among the three types of comparisons. The gaps were slightly over one half of one SD and decreased to around 0.30 SD after regression adjustments.

Among the three cognitive domains, gaps in the *applying* and *reasoning* domains were generally larger than those in the *knowing* domain except the Hispanic–White gaps, and this pattern was more pronounced in the Black–White gaps. In Figure 3d, the gap in the *knowing* domain was 0.61 SD, whereas it was 0.70 SD in the *applying* domain and 0.66 SD in the *reasoning* domain.

Associations Between Instructional Practices and Test Scores in the Reasoning Domain

Table 2 presents estimation results on the instructional practices as well as instructional time and content coverage from Equation (2). It also reports results for the *knowing* and *applying* domains as references.

Across the three cognitive domains, engaging instructional practices, same ability group work, and working on problems for which there is no immediately obvious method of solution were positively associated with the test scores.

Particularly, the same ability group work had a strong, positive relationship with the test scores. Students who worked in the same ability groups in every or almost every lesson scored 0.85 SD higher than students who never worked in the same ability groups. In sharp contrast, mixed ability group work was negatively correlated with the test scores. Students whose teachers used mixed ability group work in every or almost every lesson scored 0.64 SD lower in the *reasoning* domain than students who never worked in mixed ability groups.

The coefficient sizes on engaging instructional practices were much smaller than those on the other instructional practices. Since the variable was composed of key instructional practices that underlie many instructional approaches theoretically linked to HOTS, some of the practices might have been positively correlated, whereas the other practices could have been negatively correlated. This pattern might have attenuated the overall positive relationship. The result could also have resulted from the poor quality of the instructional practices, as the data on the practices were collected based on the *frequencies* of the practices, not the quality. Finally, it may simply mean that some students may learn

TABLE 2

Relationships Between Instructional Practices and Test Scores in the Reasoning Domain

	Knowing	Applying	Reasoning
	Model 1	Model 2	Model 3
Instructional practices			
Engaging instructional practices	0.07*** (0.01)	0.06*** (0.02)	0.05*** (0.01)
Small group work—same ability group			
Some lessons	−0.21*** (0.05)	−0.17*** (0.05)	−0.18*** (0.04)
Almost half the lessons	0.08 (0.05)	0.09* (0.05)	0.09** (0.04)
Every or almost every lesson	0.87*** (0.03)	0.78*** (0.06)	0.85*** (0.04)
Small group work—mixed ability group			
Some lessons	−0.13* (0.07)	−0.15 (0.09)	−0.09 (0.07)
Almost half the lessons	−0.68*** (0.07)	−0.64*** (0.08)	−0.59*** (0.06)
Every or almost every lesson	−0.73*** (0.05)	−0.67*** (0.07)	−0.64*** (0.05)
Work on problems with teacher guidance			
Some lessons	0.40 (0.32)	0.32 (0.37)	−0.07 (0.27)
Almost half the lessons	0.92*** (0.34)	0.78** (0.39)	0.40 (0.27)
Every or almost every lesson	0.85** (0.34)	0.71* (0.38)	0.35 (0.27)
Work on problems with no immediately obvious method of solution			
Some lessons	−0.07** (0.04)	−0.05* (0.03)	−0.04 (0.03)
Almost half the lessons	−0.13*** (0.03)	−0.13*** (0.03)	−0.11** (0.05)
Every or almost every lesson	0.46*** (0.03)	0.45*** (0.04)	0.49*** (0.05)
Instructional time			
Total hours spent on math instruction per week	−0.04** (0.02)	−0.04*** (0.01)	−0.04*** (0.02)
Content coverage			
Percentage of content in the TIMSS assessment taught by this year	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)
Constant	−2.19*** (0.54)	−2.40*** (0.58)	−1.95*** (0.39)
Student controls	X	X	X
Teacher controls	X	X	X
School fixed effects	X	X	X
Observations	6,310	6,310	6,310

Note. Mathematics teacher weights (linear transformation of total student weights) were used for all models. All plausible values were standardized. Standard errors were estimated by the Jackknife repeated replicate sampling variance estimation method combined with an unconditional approach (West et al., 2008). The unconditional approach uses the entire sample to estimate the standard errors. Sample sizes were rounded to the nearest ten due to National Center for Education Statistics nondisclosure policies.

* $p < .10$. ** $p < .05$. *** $p < .01$.

Source. U.S. Department of Education, National Center for Education Statistics, Trends in International Mathematics and Science Study 2015.

more from direct instruction, or a combination of direct instruction and student-centered instruction than student-centered instruction alone. In direct instruction, teachers carefully sequence instructions in the appropriate logical order with well-designed examples to help students make the correct inference for a new concept (Stockard et al., 2018). I explored the first possibility and reported results in the online supplement. I found that the hypothesis holds. I could not examine the other two possibilities due to a lack of data.

Another notable pattern is that the coefficient sizes on each of the instructional practices tended to be similar across the three domains. This pattern appears to suggest that HOTS may be content-specific, rather than generic. This means that engaging students in HOT requires content knowledge as well as application skills, resulting in similar coefficient sizes (for the debate about subject specificity of HOTS or critical

thinking, see the work by Bailin et al. [1999], Ennis [1987, 1989], Facione [1990], and McPeck [1981].

Outside of the instructional practices, instructional time was negatively associated with the test scores, whereas content coverage was positively related to the test scores. The negative association is counterintuitive; yet, as I explored the relationship by student subgroup in the next subsection, it appears to suggest that this might have resulted from academic tracking.

Subgroup Analysis of the Associations

In this subsection, I analyzed the relationships between the instructional practices and the test scores in the *reasoning* domain for each student subgroup to identify instructional practices that are associated with the test scores; yet,

the discussion focuses on the lowest SES students, and Black and Hispanic students. Table 3 presents the estimation results on each student subgroup from Equation 2. I reported estimation results for the *knowing* and *applying* domains in Appendix Tables 5 and 6.

The table shows that the same ability group work was strongly, positively associated with the test scores across all subgroups but Asian students, particularly the lowest SES students and Hispanic students. For example, among the lowest SES students, students who worked in the same ability groups in every or almost every lesson performed 1.76 SD higher than those who never worked in the same ability groups. The performance difference was 1.57 SD for Hispanic students.

The coefficients on the other instructional practices varied from subgroup to subgroup. For example, Black and Hispanic students who worked on problems with teacher guidance in every or almost every lesson had higher test scores than those who never did it; yet, I did not observe such a relationship among the lowest SES students. On the other hand, the lowest SES students who were asked to work on problems for which there is no immediately obvious method of solution in every or almost every lesson scored higher than those who never worked on them; such a relationship was not observed among Hispanic students. For Black students, the relationship was negative. Similarly, mixed ability group work exhibited a negative relationship among Black students but no relationship among the lowest SES students or Hispanic students. I observed similar patterns for engaging instructional practices.

Instructional time was found uncorrelated with the test scores among the lowest SES students and Black students. On the other hand, more instructional time was negatively related to the test scores among Hispanic students. A 1-hour increase in the instructional time was associated with a decline of 0.20 SD in the test score. Content coverage was positively associated with the test scores among the lowest SES students and Hispanic students but not among Black students.

Discussions and Conclusion

The current literature provides a great amount of evidence on test score gaps among student subgroups. However, most of them have focused on subject-level test score gaps, and little attention has been paid to test score gaps in HOTS. This study contributes to the current literature by examining test score gaps in the *reasoning* domain in mathematics and exploring instructional practices that may be positively associated with the test scores.

I found wide test score gaps in the *reasoning* domain by SES status and race/ethnicity, even after regression adjustments. The gaps were particularly large between Black and White students, ranging from 0.61 SD to 0.70 SD. Although the direct comparison is not possible due to differences in the methods and grades, these regression-adjusted test score gaps appear to be larger than what previous studies found at

the subject level (e.g., Clotfelter et al., 2009; Fryer & Levitt, 2004, 2006). Given the rising demand for HOTS in labor markets, this pattern is concerning and suggests that policy makers and practitioners should elaborate curricula and instructional practices to narrow the gaps.

I explored some instructional practices that prior work suggests may be associated with HOTS. I found that the same ability group work was strongly, positively associated with the test scores for all student subgroups. In contrast, mixed ability group work was negatively associated with the test scores for Black students. No relationship was found among the lowest SES students and Hispanic students. This finding appears to disagree with what researchers suggest regarding ability grouping. Yet, controversy exists regarding the efficacy of same and mixed ability grouping. A recent meta-analysis found that within-class ability grouping has at least small, positive, and significant impacts on student performance (Steenbergen-Hu et al., 2016). Although I cannot explore the mechanism due to a lack of data, Slavin (2011) argues that the type of group goals is important in group work. For example, if the teacher requires a single group product instead of individual products, more able students may do most of the group work. In this group work structure, the same ability group work may have a positive association with the test scores, but mixed ability group work could have a negative relationship. Further exploration with data on the type of group work is necessary to conclude the relationships.

Working on problems (individually or with peers) with teacher guidance was also found positively associated with the test scores among Black and Hispanic students. The TIMSS data do not provide information about what kind of and how teacher guidance was provided to the students; yet, the relationship may still underscore the importance of teacher guidance or scaffolding, which many studies found is positively correlated with student collaboration (e.g., van Leeuwen & Janssen, 2019).

In contrast, working on problems where there is no immediately obvious method of solution was found positively correlated with the test scores among all subgroups but Black and Hispanic students. In particular, the relationship was negative among Black students. Given that working on problems with teacher guidance was positively correlated with HOTS among Black and Hispanic students, this relationship may turn positive when enough scaffolding and teacher guidance are provided.

Although the focus of the second part of the analysis was on instructional practices, instructional time and content coverage are also important factors in student learning. Instructional time was found negatively correlated with the test scores for the all student sample. By student subgroup, it exhibited a positive relationship among the highest SES students and White students, whereas it had a negative relationship among Hispanic students. Although less precisely estimated, the relationship among Black students was also negative. These results may be explained by academic tracking, given that Black and

TABLE 3

Subgroup Analysis of the Relationships

	SES Q1	SES Q5	Black	Hispanic	White	Asian
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Instructional practices						
Engaging instructional practice	-0.05 (0.06)	-0.36* (0.21)	-0.06 (0.18)	-0.20*** (0.03)	-0.12** (0.06)	-1.22** (0.50)
Small group work—same ability group*						
Some lessons	0.39*** (0.10)	-0.26 (0.25)	-0.72* (0.42)	0.17*** (0.06)	-0.08 (0.08)	-4.16*** (0.87)
Almost half the lessons	0.44*** (0.08)	0.15 (0.16)	-0.42 (0.84)	0.28*** (0.03)	-0.01 (0.07)	-1.48*** (0.28)
Every or almost every lesson	1.76*** (0.18)	1.19*** (0.19)	0.61** (0.29)	1.57*** (0.07)	0.94*** (0.04)	-5.65** (2.20)
Small group work—mixed ability group*						
Some lessons	-0.15 (0.33)	0.11 (0.20)	-1.19** (0.59)	0.21 (0.19)	-0.24** (0.10)	1.68*** (0.43)
Almost half the lessons	-0.58 (0.36)	-0.21 (0.18)	-1.26 (0.86)	-0.38** (0.17)	-0.51*** (0.10)	-3.13** (1.43)
Every or almost every lesson	-0.50 (0.44)	-0.42** (0.19)	-2.02*** (0.66)	-0.35 (0.22)	-0.58*** (0.03)	0.70 (0.64)
Work on problems with teacher guidance*						
Some lessons	-0.77 (0.81)	Reference	Reference	Reference	-0.15 (0.31)	Reference
Almost half the lessons	-0.17 (0.82)	0.73*** (0.17)	0.11 (0.44)	0.47*** (0.15)	0.44 (0.36)	4.11*** (1.19)
Every or almost every lesson	-0.11 (0.84)	0.81*** (0.06)	0.70* (0.40)	0.53*** (0.12)	0.50 (0.35)	0.01 (0.61)
Work problems with no immediately obvious method of solution*						
Some lessons	-0.19* (0.11)	-0.19 (0.32)	-0.36* (0.21)	-0.10*** (0.02)	0.10** (0.05)	0.03 (0.51)
Almost half the lessons	-0.15 (0.13)	-0.26 (0.34)	-0.26 (0.34)	0.11 (0.17)	-0.01 (0.07)	4.38** (1.89)
Every or almost every lesson	0.57*** (0.08)	1.19*** (0.37)	-1.81** (0.81)	0.30 (0.24)	0.68*** (0.10)	5.17** (2.14)
Instructional time						
Total hours spent on math instruction per week	0.02 (0.03)	0.12* (0.06)	-0.07 (0.07)	-0.20*** (0.03)	0.05** (0.02)	0.26 (0.22)
Content coverage						
Percentage of content in the TIMSS assessment taught by this year	0.03*** (0.00)	0.04*** (0.00)	0.01 (0.01)	0.02*** (0.00)	0.02*** (0.00)	0.19*** (0.05)
Constant	-4.90*** (0.60)	-0.64 (2.54)	-3.43** (1.49)	-2.26*** (0.72)	-2.54*** (0.39)	-14.40** (5.75)
Student controls	X	X	X	X	X	X
Teacher controls	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X
Observations	1,130	1,370	740	1,700	3,130	280

Note. Q1= first quintile; Q5 = fifth quintile; TIMSS = Trends in International Mathematics and Science Study; SES = socioeconomic status. For instructional practices with an asterisk, the reference category was *never*, unless noted otherwise in the table. Mathematics teacher weights (linear transformation of total student weights) were used for all models. All plausible values were standardized. Standard errors were estimated by the Jackknife repeated replicate sampling variance estimation method combined with an unconditional approach (West et al., 2008). The unconditional approach uses the entire sample to estimate the standard errors. Sample sizes were rounded to the nearest ten due to National Center for Education Statistics nondisclosure policies.

* $p < .10$. ** $p < .05$. *** $p < .01$.

Source. U.S. Department of Education, National Center for Education Statistics, TIMSS 2015.

Hispanic students are more likely to be placed in basic classes. That is, the relationship was positive for the highest SES students and White students perhaps because many of them were in advanced mathematics classes; the relationship was negative for Black and Hispanic students perhaps because many of them were placed in basic classes. Content coverage was consistently positively associated with the test scores among all subgroups but Black students. The magnitude of the relationships appears to be smaller than those found among the instructional practices. Contrary to the stated hypotheses earlier, these two factors did not moderate the relationships between the instructional practices and the test scores in a meaningful way (see the online supplement).

It is important to note that, although the subgroup analysis found that engaging instructional practice was generally negatively or insignificantly associated with the test scores, further investigation reported in the online supplement revealed that some practices were positively correlated with the test scores. For example, encouraging students to express their ideas in class and asking students to explain their ideas had positive associations with the test scores among the lowest SES students and Hispanic students. All of these findings were robust to the possibility that students might have been less serious about taking the TIMSS assessment because it was a low-stakes assessment (see the online supplement).

This study faces several limitations. First, although my analysis found wide test score gaps in the *reasoning* domain and identified instructional practices that were positively associated with the test scores among low-performing

subgroups, the method was generally descriptive and did not make a causal inference. Although every effort was made, the estimated coefficients are still subject to potential bias. Second, all instructional practice variables were based on the *frequencies*, not the quality. Insignificant or negative findings on some instructional practices, therefore, do not necessarily mean that they are not recommended in the classroom. Third, data on instructional practices came from teacher self-reports. Some responses may not have accurately reflected their actual practices and contained measurement error, which might have attenuated the coefficients on some variables. Finally, although not the focus of this study, instructional time did not measure how much time was spent on HOT activities, and the distribution of time across the three cognitive domains was unknown. The coefficients on instructional time as well as its moderation effects (see the online supplement) should be interpreted with caution.

Even with these limitations, this study makes a significant contribution to the current literature and provides the direction for future research. The study benefits from future research that collects more rich data on instructional practices that may be directly related to the instructional practices described earlier. It also benefits from studies that examine test score gaps in HOTS in reading and science in the same and different grades. Finally, research on whether and how HOTS acquired at K–12 systems leads to success in later adult outcomes would help practitioners make seamless the alignment in educational standards and goals from PreK–12 to higher education to career.

APPENDIX TABLE 1

Description of Bloom's Revised Taxonomy

Categories & Cognitive Processes	Alternative Names	Definitions and Examples
Analyze: Break material into its constituent parts and determine how the parts relate to one another and to an overall structure or purpose		
Differentiating	Discriminating, distinguishing, focusing, selecting	Distinguishing relevant from irrelevant parts or important from unimportant parts of presented material (e.g., Distinguish between relevant and irrelevant numbers in a mathematical word problem)
Organizing	Finding, coherence, integrating, outlining, parsing, structuring	Determine how elements fit or function within a structure (e.g., Structure evidence in a historical description into evidence for and against a particular historical explanation)
Attributing	Deconstructing	Determine a point of view, bias, values, or intent underlying presented material (e.g., Determine the point of view of the author of an essay in terms of his or her political perspective)
Evaluate: Make judgments based on criteria and standards		
Checking	Coordinating, detecting, monitoring, testing	Detecting inconsistencies or fallacies within a process or product; determining whether a process or product has internal consistency; detecting the effectiveness of a procedure as it is being implemented (e.g., Determine if a scientist's conclusions follow from observed data)
Critiquing	Judging	Detecting inconsistencies between a product and external criteria; determining whether a product has external consistency; detecting the appropriateness of a procedure for a given problem (e.g., Judge which of two methods is the best way to solve a given problem)
Create: Put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure		
Generating	Hypothesizing	Coming up with alternative hypotheses based on criteria (e.g., Generate hypotheses to account for an observed phenomenon)
Planning	Designing	Devising a procedure for accomplishing some task (e.g., Plan a research paper on a given historical topic)
Producing	Constructing	Inventing a product (e.g., Build habitats for a specific purpose)

SOURCE: Anderson, L. W. (Ed.), Krathwohl, D. R. (Ed.), Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). Longman.

APPENDIX TABLE 2

Description of TIMSS Cognitive Domain

Panel A: Knowing

Recall	Recall definitions, terminology, number properties, units of measurement, geometric properties, and notation (e.g., $a \times b = ab$, $a + a + a = 3a$).
Recognize	Recognize numbers, expressions, quantities, and shapes. Recognize entities that are mathematically equivalent (e.g., equivalent familiar fractions, decimals, and percents; different orientations of simple geometric figures).
Classify/order	Classify numbers, expressions, quantities, and shapes by common properties.
Compute	Carry out algorithmic procedures for $+$, $-$, \times , \div , or a combination of these with whole numbers, fractions, decimals, and integers. Carry out straightforward algebraic procedures.
Retrieve	Retrieve information from graphs, tables, texts, or other sources.
Measure	Use measuring instruments; and choose appropriate units of measurement.

Panel B: Applying

Determine	Determine efficient/appropriate operations, strategies, and tools for solving problems for which there are commonly used methods of solution.
Represent/Model	Display data in tables or graphs; create equations, inequalities, geometric figures, or diagrams that model problem situations; and generate equivalent representations for a given mathematical entity or relationship.
Implement	Implement strategies and operations to solve problems involving familiar mathematical concepts and procedures.

Panel C: Reasoning

Analyze	Determine, describe, or use relationships among numbers, expressions, quantities, and shapes.
Integrate/synthesize	Link different elements of knowledge, related representations, and procedures to solve problems.
Evaluate	Evaluate alternative problem solving strategies and solutions.
Draw conclusions	Make valid inferences on the basis of information and evidence.
Generalize	Make statements that represent relationships in more general and more widely applicable terms.
Justify	Provide mathematical arguments to support a strategy or solution.

SOURCE: Grønmo, L. S., Lindquist, M., Arora, A., & Mullis, I. V. S. (2013). TIMSS 2015 mathematics framework. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 11-27). TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College; International Association for the Evaluation of Educational Achievement (IEA).

APPENDIX TABLE 3

*Survey Items Used for the Socioeconomic Status Variable***Student-level survey items**

Number of books (0-10 books, 11-25 books, 26-100 books, 101-200 books more than 200 books)

About how many books are there in your home?

Possessions (Yes or No)

Computer (do not include PlayStation®, GameCube®, Xbox®, or other TV/video game systems)

Computer or tablet that is shared with other people at home

Study desk/table for your use

Your own room

Your own cell

Internet connection

PlayStation®, GameCube®, Xbox®, or other TV/video game systems

VCR or DVD player

Activities outside of school (Yes or No)

Do you play on a sports team outside of school?

Do you often play a musical instrument outside of school?

Are you studying something in a class outside of school?

Do you belong to a club outside of school?

School-level SES measure from CCD

Percent of students eligible for the federal free/reduced lunch program

SOURCE: U. S. Department of Education, National Center for Education Statistics, Trends in International Mathematics and Science Study (TIMSS) 2015.

APPENDIX TABLE 4

Survey Item Used for Engaging Instructional Practice

	Mean	SD
Engaging instructional practices (1= <i>never</i> ; 4= <i>every or almost every lesson</i>)		
Relate the lesson to students' daily lives	2.99	0.84
Ask students to explain their answers	3.70	0.58
Ask students to complete challenging exercises that require them to go beyond the instruction	3.03	0.83
Encourage classroom discussions among students	3.40	0.80
Link new content to students' prior knowledge	3.71	0.60
Ask students to decide their own problem solving procedures	3.12	0.79
Encourage students to express their ideas in class	3.54	0.73

Note. These statistics were estimated based on the analytic sample used for the subsequent analyses (N=6310). Mathematics teacher weights (linear transformation of total student weights) were used to estimate means and standard deviations. Although not reported, standard errors were estimated by the Jackknife Repeated Replicate sampling variance estimation method. By survey design, statistics on teachers do not represent a teacher population.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Trends in International Mathematics and Science Study (TIMSS) 2015.

APPENDIX TABLE 5

Subgroup Analysis of the Relationships – Knowing

	SES 1st Quintile	SES 5th Quintile	Black	Hispanic	White	Asian
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Instructional strategies to improve HOTS						
Engaging instructional practice	-0.07 (0.06)	-0.34 (0.24)	-0.09 (0.14)	-0.26*** (0.04)	-0.10* (0.05)	-1.18* (0.68)
Small group work - same ability group*						
Some lessons	0.31*** (0.11)	-0.46** (0.22)	-0.84*** (0.31)	-0.12* (0.06)	-0.08 (0.07)	-4.92*** (0.92)
Almost half the lessons	0.37*** (0.14)	-0.02 (0.08)	-0.81 (0.63)	0.34*** (0.03)	-0.02 (0.10)	-1.26 (0.87)
Every or almost every lesson	1.72*** (0.25)	1.10*** (0.16)	0.43** (0.19)	1.60*** (0.13)	1.01*** (0.06)	-7.29*** (2.65)
Small group work - mixed ability group*						
Some lessons	0.02 (0.25)	0.07 (0.16)	-1.19** (0.60)	0.07 (0.37)	-0.27** (0.11)	1.38*** (0.42)
Almost half the lessons	-0.48 (0.31)	-0.35*** (0.13)	-0.92 (0.76)	-0.58* (0.31)	-0.59*** (0.13)	-4.48*** (1.40)
Every or almost every lesson	-0.44 (0.33)	-0.51** (0.22)	-1.89*** (0.67)	-0.50 (0.36)	-0.66*** (0.11)	0.01 (0.56)
Work problems with teacher guidance*						
Some lessons	-0.45 (0.64)	Reference	Reference	Reference	0.04 (0.35)	Reference
Almost half the lessons	0.12 (0.64)	0.86*** (0.19)	-0.27 (0.18)	0.49*** (0.12)	0.69** (0.35)	5.45*** (1.34)
Every or almost every lesson	0.16 (0.62)	0.94*** (0.07)	0.30* (0.16)	0.55*** (0.11)	0.73** (0.35)	-0.13 (0.59)
Work problems with no immediately obvious method of solution*						
Some lessons	-0.26* (0.13)	-0.19 (0.35)	-0.23 (0.20)	-0.12*** (0.03)	0.08 (0.07)	0.39 (0.52)
Almost half the lessons	-0.08 (0.14)	-0.26 (0.34)	-0.12 (0.24)	0.15 (0.14)	-0.06 (0.06)	6.14*** (2.10)
Every or almost every lesson	0.50*** (0.15)	1.16*** (0.43)	-1.82*** (0.64)	0.34 (0.23)	0.69*** (0.10)	7.50*** (2.21)
Instructional time						
Total hours spent on math instruction per week	0.02 (0.03)	0.13* (0.08)	-0.09* (0.05)	-0.21*** (0.04)	0.05* (0.03)	0.47** (0.20)

(continued)

APPENDIX TABLE 5 (continued)

	SES 1st Quintile	SES 5th Quintile	Black	Hispanic	White	Asian
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Content coverage						
Percent of content in the TIMSS assessment taught by this year	0.03*** (0.00)	0.04*** (0.01)	0.01 (0.01)	0.02*** (0.00)	0.03*** (0.00)	0.23*** (0.06)
Constant	-4.99*** (0.88)	-0.16 (2.74)	-2.93** (1.16)	-1.69*** (0.50)	-2.79*** (0.61)	-17.31*** (6.22)
Student controls	X	X	X	X	X	X
Teacher controls	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X
Observations	1,130	1,370	740	1,700	3,130	280

Note. For instructional practices with an asterisk, the reference category was *never*, unless noted otherwise in the table. Mathematics teacher weights (linear transformation of total student weights) were used for all models. All plausible values were standardized. Standard errors were estimated by the Jackknife Repeated Replicate sampling variance estimation method combined with an unconditional approach (West et al., 2008). The unconditional approach uses the entire sample to estimate the standard errors. Sample sizes were rounded to the nearest ten due to NCES non-disclosure policies.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Trends in International Mathematics and Science Study (TIMSS) 2015.

APPENDIX TABLE 6

Subgroup Analysis of the Relationships – Applying

	SES 1st Quintile	SES 5th Quintile	Black	Hispanic	White	Asian
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Instructional strategies to improve HOTS						
Engaging instructional practice	-0.06 (0.05)	-0.37* (0.20)	-0.17 (0.12)	-0.22*** (0.04)	-0.12** (0.05)	-1.04* (0.63)
Small group work - same ability group*						
Some lessons	0.44*** (0.07)	-0.34 (0.23)	-0.99** (0.43)	-0.11 (0.09)	-0.09 (0.06)	-3.72*** (0.86)
Almost half the lessons	0.46*** (0.11)	0.09 (0.13)	-0.88 (0.76)	0.32*** (0.03)	-0.03 (0.09)	-0.95 (0.88)
Every or almost every lesson	1.72*** (0.24)	1.00*** (0.14)	0.25 (0.21)	1.47*** (0.07)	0.86*** (0.09)	-4.79** (2.37)
Small group work - mixed ability group*						
Some lessons	-0.37 (0.27)	0.02 (0.20)	-1.06** (0.47)	-0.01 (0.26)	-0.24** (0.10)	1.21*** (0.13)
Almost half the lessons	-0.79** (0.36)	-0.29* (0.16)	-0.90 (0.64)	-0.60*** (0.23)	-0.52*** (0.10)	-2.91** (1.21)
Every or almost every lesson	-0.73** (0.36)	-0.45*** (0.17)	-1.84*** (0.62)	-0.53* (0.29)	-0.56*** (0.10)	0.39 (0.57)
Work problems with teacher guidance*						
Some lessons	-0.18 (0.81)	Reference	Reference	Reference	0.10 (0.43)	Reference
Almost half the lessons	0.39 (0.87)	0.68*** (0.18)	-0.26 (0.35)	0.45** (0.19)	0.69 (0.44)	3.22** (1.31)
Every or almost every lesson	0.44 (0.86)	0.76*** (0.08)	0.38 (0.25)	0.50*** (0.15)	0.73* (0.44)	-0.22 (0.49)
Work problems with no immediately obvious method of solution*						
Some lessons	-0.21* (0.11)	-0.18 (0.30)	-0.31* (0.17)	-0.12** (0.05)	0.12* (0.07)	-0.12 (0.63)
Almost half the lessons	-0.17 (0.14)	-0.22 (0.28)	-0.26 (0.16)	0.11 (0.15)	-0.02 (0.05)	3.62* (1.96)
Every or almost every lesson	0.47*** (0.13)	1.20*** (0.34)	-2.02*** (0.73)	0.38* (0.21)	0.65*** (0.10)	4.34** (2.15)

(continued)

APPENDIX TABLE 6 (continued)

	SES 1st Quintile	SES 5th Quintile	Black	Hispanic	White	Asian
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Instructional time						
Total hours spent on math instruction per week	0.02 (0.04)	0.12* (0.07)	-0.10** (0.04)	-0.19*** (0.03)	0.06*** (0.02)	0.23 (0.26)
Content coverage						
Percent of content in the TIMSS assessment taught by this year	0.03*** (0.00)	0.04*** (0.00)	0.01 (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.15*** (0.05)
Constant	-5.85*** (0.98)	-0.43 (2.34)	-3.40*** (1.30)	-2.09*** (0.62)	-2.97*** (0.59)	-9.56* (5.68)
Student controls	X	X	X	X	X	X
Teacher controls	X	X	X	X	X	X
School fixed effects	X	X	X	X	X	X
Observations	1,130	1,370	740	1,700	3,130	280

Note. For instructional practices with an asterisk, the reference category was *never*, unless noted otherwise in the table. Mathematics teacher weights (linear transformation of total student weights) were used for all models. All plausible values were standardized. Standard errors were estimated by the Jackknife Repeated Replicate sampling variance estimation method combined with an unconditional approach (West et al., 2008). The unconditional approach uses the entire sample to estimate the standard errors. Sample sizes were rounded to the nearest ten due to NCES non-disclosure policies.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Trends in International Mathematics and Science Study (TIMSS) 2015.

Author Note

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. There is no single definition of what constitutes 21st-century skills. Although the framework is similar, many organizations, researchers, practitioners, and companies also define 21st-century skills differently.

2. Researchers debate whether critical thinking or HOTS in the current context is generic or content specific. For those interested, refer to the work by Bailin et al. (1999), Ennis (1987, 1989), Facione (1990), and McPeck (1981).

3. Kurz and his colleagues argue that opportunity to learn is a multidimensional concept consisting of instructional time, content coverage, and quality of instruction (Elliott & Bartlett, 2016; Kurz, 2011). This broader definition is more frequently used in educational effectiveness research.

4. I obtained restricted-use U.S. TIMSS data from the Institute of Education Sciences, U.S. Department of Education.

5. A total of 18 states in the U.S. participated as benchmarking states in TIMSS since 1995 (U.S. Department of Education, n.d.).

6. For details of the targeted percentages of assessment items devoted to the content and cognitive domains, and the definition of each process, refer to a report by Mullis et al. (2009).

7. The restricted-use data also included the same information. However, I chose to use the data from the CCD because they had fewer missing data than the former data.

8. Sample sizes were rounded to the nearest 10 due to the National Center for Education Statistics nondisclosure policies.

9. The TIMSS assessments prior to the 2007 administration used different numbers of booklets.

10. Test scores in the *reasoning* domain could have been influenced by students' knowledge. For example, a student may not have been able to answer a problem classified into the *reasoning* domain because they did not remember a formula or a theorem, although they had adequate HOTS to answer the problem. In this sense, the test scores were underestimated to a certain degree. Similarly, positive relationships between some instructional practices and the test scores in the subsequent analyses may have been driven by improving students' knowledge, rather than their HOTS. In this sense, the estimated coefficients may have been overestimated to a certain degree.

11. I also created an SES variable that incorporates parents' highest education level (*bsdgedup*) and reanalyzed achievement gaps. The correlation between the SES variable with and without parents' highest education level was 0.98. Findings are similar.

12. Note that this SES variable was not a perfect measure of SES, as its components did not include direct measures of family income, parental educational attainment, or parental occupational status. Yet, as noted in Note 11, the correlation between the SES variable with and without parents' highest education level was 0.98. As a validity check, I calculated a correlation between the percentage of students eligible for the federal free/reduced lunch program and the school-level average of the SES variable. It was -0.92, suggesting that the SES variable closely captures an aspect of family income.

13. An exploratory factor analysis revealed a clear, single factor, which is almost perfectly correlated with the item-mean variable ($\rho = 0.9978$). For ease of interpretation, I used the item-mean variable.

14. Asking students to work problems with teacher guidance could be also considered as direction instruction, depending on how teacher guidance is interpreted. I interpreted it as scaffolding that the teacher provided for their students.

15. The wording of the two response options, *mostly taught this year* and *not yet taught or just introduced*, suggests that the indicator variable may contain a measurement error. For example, *mostly taught this year* may mean that 75% of the topic was taught instead of 100%. Similarly, *not yet taught or just introduced* may mean that 25% of the topic was just taught. To check the robustness of the results to this possible issue, I created two different content coverage variables. One was based on the set of 20 variables that take a value of unity if the topic was taught before this year, a value of 0.75 if the topic was mostly taught this year, and a value of 0 if the topic was not yet taught or just introduced. The other variable was based on the set of 20 variables that take a value of unity if the topic was taught before this year or mostly taught this year and a value of 0.25 if the topic was not yet taught or just introduced. I reestimated Equation (2) using these variables separately and found similar results (results not reported here).

16. A TIMSS technical report, *Methods and Procedures in TIMSS 2015* (M. O. Martin et al., 2016), provides a detailed description of how this variable was created. It also explains teachers' job satisfaction used in this study.

17. Each of the five plausible values was standardized.

18. There are generally three ways to measure test score gaps between student subgroups: the difference in mean test scores, the difference in standardized mean test scores, and metric-free measures of the gaps (Ho & Haertel, 2006; Reardon & Galindo, 2009; Reardon & Robinson, 2007) or, more generally, distributional measures (e.g., Handcock & Morris, 1998, 1999). Ho and Haertel (2006), Reardon and Galindo (2009), Reardon and Robinson (2007), and Handcock and Morris (1998, 1999) discuss limitations that each type of measure has.

19. Enough variation exists in instructional practices within schools to estimate the associations.

20. The model was estimated using mathematics teacher weights, which are a linear transformation of total student weights (Rutkowski et al., 2010). Standard errors were estimated using the Jackknife repeated replication variance estimation method.

21. About 62% of the schools in the sample used student achievement to assign eight-grade students to mathematics classes. However, because the focus of the current investigation was on the relationships that exist across all classroom types, rather than within academic track classes, I did not control for academic tracking.

References

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85(2), 275–314. <https://doi.org/10.3102/0034654314551063>
- Allison, P. D. (2002). *Missing data series: Quantitative applications in the social sciences*. Sage.
- Altintas, E. (2016). The widening education gap in developmental child care activities in the United States, 1965–2013. *Journal of Marriage and Family*, 78(1), 26–42. <https://doi.org/10.1111/jomf.12254>
- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Rath, J., & Wittrock, M. C. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete ed.). Longman.
- Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258–267. <https://doi.org/10.3102/0013189X07306523>
- Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies*, 31(3), 285–302. <https://dx.doi.org/10.1080/002202799183133>
- Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Longmans.
- Boaler, J. (2002). Learning from teaching: Exploring the relationship between reform curriculum and equity. *Journal for Research in Mathematics Education*, 33(4), 239–258. <https://doi.org/10.2307/749740>
- Bonney, C. R., & Sternberg, R. J. (2011). Learning to think critically. In R. E. Mayer, & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 166–196). Routledge.
- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Harvard University Press.
- Bronfenbrenner, U. (1986). Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, 22(6), 723–742. <https://doi.org/10.1037/0012-1649.22.6.723>
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Association for Supervision & Curriculum Development.
- Brunner, J. L. (2014). *What factors help or hinder the achievement of low SES students? An international comparison using TIMSS 2011 8th grade science data* [Doctoral dissertation]. Michigan State University.
- Burris, C. C., & Welner, K. G. (2005). Closing the achievement gap by detracking. *Phi Delta Kappan*, 86(8), 594–598. <https://doi.org/10.1177/003172170508600808>
- Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, 64(8), 723–733. https://doi.org/10.1007/978-1-4419-1428-6_980
- Castro, A. J., Kelly, J., & Shih, M. (2010). Resilience strategies for new teachers in high-needs areas. *Teaching and Teacher Education*, 26(3), 622–629. <https://doi.org/10.1016/j.tate.2009.09.010>
- Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, 24(4), 377–392. <https://doi.org/10.1016/j.econedurev.2004.06.008>
- Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2009). The academic achievement gap in Grades 3 to 8. *Review of Economics and Statistics*, 91(2), 398–419. <https://doi.org/10.1162/rest.91.2.398>
- Clotfelter, C., Ladd, H. F., Vigdor, J., & Wheeler, J. (2007). High-poverty schools and the distribution of teachers and principals. *North Carolina Law Review*, 85, 1345–1379. <https://doi.org/10.1037/e722752011-001>
- Common Core State Standards Initiative. (n.d.). *Standards for Mathematical Practice*. <http://www.corestandards.org/Math/Practice/>
- Cowan, C. D., Hauser, R. M., Kominski, R. A., Levin, H. M., Lucas, S. R., Mogan, S. L., Spencer, M. B., & Chapman, C. (2012). *Improving the measurement of socioeconomic status for the National Assessment of Educational Progress: A theoretical foundation* (Recommendations to the National Center for

- Education Statistics). https://nces.ed.gov/nationsreportcard/pdf/researchcenter/Socioeconomic_Factors.pdf
- Crosnoe, R., Morrison, F., Burchinal, M., Pianta, R., Keating, D., Friedman, S. L., & Network, E. C. C. R. (2010). Instruction, teacher-student relations, and math achievement trajectories in elementary school. *Journal of Educational Psychology, 102*(2), 407–417. <https://doi.org/10.1037/a0017762>
- Curran, C. F. (2017). Income-based disparities in early elementary school science achievement. *Elementary School Journal, 118*(2), 207–231. <https://doi.org/10.1086/694218>
- Curran, C. F., & Kellogg, A. T. (2016). Understanding science achievement gaps by race/ethnicity and gender in kindergarten and first grade. *Educational Researcher, 45*(5), 273–282. <https://doi.org/10.3102/0013189X16656611>
- Darling-Hammond, L. (2001). Inequality in teaching and schooling: How opportunity is rationed to students of color in America. In *The right thing to do, the smart thing to do: Enhancing diversity in health professions—Summary of the symposium on diversity in health professions in honor of Herbert W. Nickens, M.D.* (pp. 208–233). National Academies Press, Institute of Medicine. <https://doi.org/10.17226/10186>
- de Jong, T., & Lazonder, A. W. (2014). The guided discovery learning principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 371–390). Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369>
- Desimone, L. M., & Long, D. (2010). Teacher effects and the achievement gap: Do teacher and teaching quality influence the achievement gap between black and white and high- and low-SES students in the early grades? *Teachers College Record, 112*(12), 3024–3073.
- Dias-Lacy, S. L., & Guirguis, R. V. (2017). Challenges for new teachers and ways of coping with them. *Journal of Education and Learning, 6*(3), 265–272. <https://doi.org/10.5539/jel.v6n3p265>
- Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. In R. Murnane, & G. J. Duncan (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 47–69). Russell Sage Foundation.
- Elliott, S., & Bartlett, B. (2016). *Opportunity to learn*. Oxford Handbooks Online. <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199935291.001.0001/oxfordhb-9780199935291-e-70>
- Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and skills. In J. Baron, & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9–26). W. H. Freeman.
- Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*(3), 4–10. <https://doi.org/10.3102/0013189X018003004>
- Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice, 32*(3), 179–186. <https://doi.org/10.1080/00405849309543594>
- Etkina, E., Karelina, A., & Ruibal-Villasenor, M. (2008). How long does it take? A study of student acquisition of scientific abilities. *Physical Review Physics Education Research, 4*(2), 020108. <https://doi.org/10.1103/PhysRevSTPER.4.020108>
- Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations*. American Philosophical Association.
- Fayer, S., Lacey, A., & Watson, A. (2017). *STEM occupations: Past, present, and future*. U.S. Bureau of Labor Statistics.
- Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16*(2), 234–248. <https://doi.org/10.1111/desc.12019>
- Floden, R. E. (2002). The measurement of opportunity to learn. In *Methodological advances in cross-national surveys of educational achievement* (pp. 231–266). National Research Council & National Academies Press. <https://doi.org/10.17226/10322>
- Frausel, R. R., Silvey, C., Freeman, C., Dowling, N., Richland, L. E., Levine, S. C., Raudenbush, S., & Goldin-Meadow, S. (2020). The origins of higher-order thinking lie in children's spontaneous talk across the pre-school years. *Cognition, 200*(2020), 1–24. <https://doi.org/10.1016/j.cognition.2020.104274>
- 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>
- Fryer, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review, 8*(2), 249–281. <https://doi.org/10.1093/aler/ahl003>
- Gamoran, A. (1987). The stratification of high school learning opportunities. *Sociology of Education, 60*(3), 135–155. <https://doi.org/10.2307/2112271>
- Gillies, R. M. (2011). Promoting thinking, problem-solving and reasoning during small group discussions. *Teachers and Teaching, 17*(1), 73–89. <https://doi.org/10.1080/13540602.2011.538498>
- Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher, 44*(5), 293–307. <https://doi.org/10.3102/0013189X15592622>
- Grønmo, L. S., Lindquist, M., Arora, A., & Mullis, I. V. S. (2013). TIMSS 2015 mathematics framework. In I. V. S. Mullis, & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 11–27). Boston College, International Association for the Evaluation of Educational Achievement.
- Guskey, T. R. (2007). Closing achievement gaps: Revisiting Benjamin S. Bloom's "Learning for Mastery." *Journal of Advanced Academics, 19*(1), 8–31. <https://doi.org/10.4219/jaa-2007-704>
- Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*(4), 449–455. <https://doi.org/10.1037/0003-066X.53.4.449>
- Handcock, M. S., & Morris, M. (1998). Relative distribution methods. *Sociological Methodology, 28*(1), 53–97. <https://doi.org/10.1111/0081-1750.00042>
- Handcock, M. S., & Morris, M. (1999). *Relative distribution methods in the social sciences*. Springer. <https://doi.org/10.1007/b97852>
- Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources, 39*(2), 326–354. <https://doi.org/10.3368/jhr.XXXIX.2.326>
- Hanushek, E. A., & Rivkin, S. G. (2006). *School quality and the black-white achievement gap* (Working Paper No. 12651). National Bureau of Economic Research. <http://www.nber.org/papers/w12651>

- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H. Brookes.
- He, Y., & Cooper, J. (2011). Struggles and strategies in teaching: Voices of five novice secondary teachers. *Teacher Education Quarterly*, 38(2), 97–116.
- 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* (NCES 2011-459). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
- Hitchcock, D. (2020). Critical thinking. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 ed.). <https://plato.stanford.edu/archives/fall2020/entries/critical-thinking/>
- Hmelo, C. E., & Ferrari, M. (1997). The problem-based learning tutorial: Cultivating higher order thinking skills. *Journal for the Education of the Gifted*, 20(4), 401–422. <https://doi.org/10.1177/016235329702000405>
- Ho, A. D., & Haertel, E. H. (2006). *Metric-free measures of test score trends and gaps with policy-relevant examples* (CSE Report No. 665). Center for the Study of Evaluation.
- Hung, W., Jonassen, D. H., & Liu, R. (2008). Problem-based learning. In M. Spector, D. Merrill, J. van Merriënboer, & M. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 485–506). Erlbaum. <https://doi.org/10.1007/978-1-4614-3185-5>
- Jencks, C., & Phillips, M. (Eds.). (1998). *The black-white test score gap*. Brookings Institute Press.
- Kalil, A. (2015). Inequality begins at home: The role of parenting in the diverging destinies of rich and poor children. In P. Amato, A. Booth, S. McHale, & J. Van Hook (Eds.), *Diverging destinies: Families in an era of increasing inequality* (pp. 63–82). Springer. <https://dx.doi.org/10.1007/978-3-319-08308-7>
- Kalil, A., Ziol-Guest, K. M., Ryan, R. M., & Markowitz, A. J. (2016). Changes in income-based gaps in parent activities with young children from 1988 to 2012. *AERA Open*, 2(3), 1–17. <https://doi.org/10.1177/2332858416653732>
- King, A. (1992). Facilitating elaborative learning through guided student-generated questioning. *Educational Psychologist*, 27(1), 111–126. https://dx.doi.org/10.1207/s15326985ep2701_8
- King, A. (2002). Structuring peer interaction to promote high-level cognitive processing. *Theory Into Practice*, 41(1), 33–39. https://doi.org/10.1207/s15430421tip4101_6
- Krathwohl, D. R. (2002). A revision of Bloom’s Taxonomy: An overview. *Theory Into Practice*, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- Kurz, A. (2011). Access to what should be taught and will be tested: Students’ opportunity to learn the intended curriculum. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *The handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy* (pp. 99–129). Springer. <https://doi.org/10.1007/978-1-4419-9356-4>
- Ladson-Billings, G. (1997). It doesn’t add up: African American students’ mathematics achievement. *Journal for Research in Mathematics Education*, 28(6), 697–708. <https://doi.org/10.2307/749638>
- Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher*, 35(7), 3–12. <https://doi.org/10.3102/0013189X035007003>
- Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, 24(1), 37–62. <https://doi.org/10.3102/01623737024001037>
- Lareau, A. (2003). *Unequal childhoods: Class, race, and family life*. University of California Press.
- Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, 86(3), 681–718. <https://doi.org/10.3102/0034654315627366>
- Lewis, A., & Smith, D. (1993). Defining higher order thinking. *Theory Into Practice*, 32(3), 131–137. <https://doi.org/10.1080/00405849309543588>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons. <https://doi.org/10.1002/9781119013563>
- Love, B. J. (2004). Brown plus 50 counter-storytelling: A critical race theory analysis of the “majoritarian achievement gap” story. *Equity & Excellence in Education*, 37(3), 227–246. <https://doi.org/10.1080/10665680490491597>
- Lu, J., Bridges, S., & Hmelo-Silver, C. E. (2014). Problem-based learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 298–318). Cambridge University Press. <https://doi.org/10.1017/CBO9781139519526>
- Luyten, H. (2017). Predictive power of OTL measures in TIMSS and PISA. In J. Scheerens (Ed.), *Opportunity to learn, curriculum alignment and test preparation: A research review* (pp. 103–119). Springer. <https://dx.doi.org/10.1007/978-3-319-43110-9>
- Maas, T., Jochim, A., & Gross, B. (2018). *Mind the gap: Will all students benefit from 21st century learning?* University of Washington.
- Marshall, J. C., & Horton, R. M. (2011). The relationship of teacher-facilitated, inquiry-based instruction to student higher-order thinking. *School Science and Mathematics*, 111(3), 93–101. <https://doi.org/10.1111/j.1949-8594.2010.00066.x>
- Martin, L. M., & Halpern, D. F. (2011). Pedagogy for developing critical thinking skills in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity*, 6(1), 1–13. <https://doi.org/10.1016/j.tsc.2010.08.002>
- Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Boston College.
- Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Boston College.
- Masek, A., & Yamin, S. (2011). The effect of problem based learning on critical thinking ability: A theoretical and empirical review. *International Review of Social Sciences and Humanities*, 2(1), 215–221. <https://dx.doi.org/10.2224/sbp.2006.34.9.1127>
- McPeck, J. (1981). *Critical thinking and education*. St. Martin’s Press. <https://doi.org/10.2307/3121218>
- Mehta, J. (2014, June 20). Deeper learning has a race problem. *Education Week*. http://blogs.edweek.org/edweek/learning_deeply/2014/06/deeper_learning_has_a_race_problem.html
- Mickelson, R. A. (2001). Subverting Swann: First- and second-generation segregation in the Charlotte-Mecklenburg schools. *American Educational Research Journal*, 38(2), 215–252. <https://doi.org/10.3102/00028312038002215>
- Miri, B., David, B., & Uri, Z. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical

- thinking. *Research in Science Education*, 37, 353–369. <https://doi.org/10.1007/s11165-006-9029-2>
- Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 Assessment Frameworks*. Boston College. <http://timssandpirls.bc.edu/timss2015/frameworks.html>
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Boston College. https://timssandpirls.bc.edu/timss2011/downloads/TIMSS2011_Frameworks.pdf
- Murphy, P. K., Firetto, C. M., Wei, L., Li, M., & Croninger, R. M. V. (2016). What really works: Optimizing classroom discussions to promote comprehension and critical-analytic thinking. *Policy Insights from the Behavioral and Brain Sciences*, 3(1), 27–35. <https://doi.org/10.1177/2372732215624215>
- National Council of Teachers of Mathematics. (2018). *Building STEM education on a sound mathematical foundation*. <https://www.nctm.org/Standards-and-Positions/Position-Statements/Building-STEM-Education-on-a-Sound-Mathematical-Foundation/>
- Noguera, P., Darling-Hamond, L., & Friedlaender, D. (2015). *Equal opportunity for deeper learning*. Jobs for the Future.
- Oakes, J. (2005). *Keeping track: How schools structure inequality* (2nd ed.). Yale University Press.
- Partnership for 21st Century Learning. (n.d.). *Framework for 21st century learning*. <http://www.p21.org/our-work/p21-framework>
- Patrick, K., Socol, A., & Morgan, I. (2020). *Inequities in advanced coursework: What's driving them and what leaders can do*. Education Trust.
- Pew Research Center. (2016). *The state of American jobs*. http://assets.pewresearch.org/wp-content/uploads/sites/3/2016/10/ST_2016.10.06_Future-of-Work_FINAL4.pdf
- Phillips, M. (2011). Parenting, time use, and disparities in academic outcomes. In R. Murnane, & G. J. Duncan (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 207–228). Russell Sage Foundation.
- Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, 36(4), 399–416. <https://doi.org/10.3102/0162373714531851>
- Quinn, D. M., Desruisseaux, T., & Nkansah-Amankra, A. (2019). "Achievement gap" language affects teachers' issue prioritization. *Educational Researcher*, 48(7), 484–487. <https://doi.org/10.3102/0013189X19863765>
- Reardon, S. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In R. Murnane, & G. J. Duncan (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 91–116). Russell Sage Foundation.
- Reardon, S., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, 46(3), 853–891. <https://doi.org/10.3102/0002831209333184>
- Reardon, S., & Robinson, J. P. (2007). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. A. Ladd, & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (1st ed., pp. 497–516). Taylor & Francis. <https://doi.org/10.4324/9780203961063>
- Richland, L. E., & Begolli, K. N. (2016). Analogy and higher order thinking: Learning mathematics as an example. *Instructional Strategy*, 3(2), 160–168. <https://doi.org/10.1177/2372732216629795>
- Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *WIREs Cognitive Science*, 6(2), 177–192. <http://doi.org/10.1002/wcs.1336>
- Rogers, J., Mirra, N., Seltzer, M., & Jun, J. (2014). *It's about time: Learning time and educational opportunity in California high schools*. University of California, IDEA.
- Rumberger, R. W., & Tran, L. (2010). State language policies, school language practices, and the English learner achievement gap. In P. Gándara, & M. Hopkins (Eds.), *Forbidden languages: English learners and restrictive language policies* (pp. 86–101). Teachers College Press.
- Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142–151. <https://doi.org/10.3102/0013189X10363170>
- Scheerens, J. (Ed.). (2017). *Opportunity to learn, curriculum alignment and test preparation: A research review*. Springer. <https://doi.org/10.1007/978-3-319-43110-9>
- Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher*, 44(7), 371–386. <https://doi.org/10.3102/0013189X15603982>
- Schmidt, W. H., Houang, R. T., Cogan, L. S., & Solorio, M. L. (2019). *Schooling across the globe: What we have learned from 60 years of mathematics and science international assessments*. Cambridge University Press. <https://doi.org/10.1017/9781316758830>
- Schraw, G., McCrudden, M. T., Lehman, S., & Hoffman, B. (2011). An overview of thinking skills. In G. Schraw, & D. R. Robinson (Eds.), *Assessment of higher order thinking skills* (pp. 19–45). Information Age.
- Slavin, R. E. (2011). Instruction based on cooperative learning. In R. E. Mayer, & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 344–360). Routledge.
- Snyder, L. G., & Snyder, M. J. (2008). Teaching critical thinking and problem solving skills. *Delta Pi Epsilon Journal*, 50(2), 90–99.
- Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K-12 students' academic achievement: Findings of two second-order meta-analysis. *Review of Educational Research*, 86(4), 849–899. <https://doi.org/10.3102/0034654316675417>
- Stockard, J., Wood, T. W., Coughlin, C., & Khoury, C. R. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, 88(4), 479–507. <https://doi.org/10.3102/0034654317751919>
- U.S. Department of Education. (n.d.). *TIMSS participating countries*. <https://nces.ed.gov/timss/participation.asp>
- van Leeuwen, A., & Janssen, J. (2019). A systematic review of teacher guidance during collaborative learning in primary and secondary education. *Educational Research Review*, 27(June), 71–89. <https://doi.org/10.1016/j.edurev.2019.02.001>

- von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In M. von Davier, & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (IERI Monograph Series, Vol. 2, pp. 9–36). Educational Testing Service.
- Warburton, E., & Torff, B. (2005). The effect of perceived learner advantages on teachers' beliefs about critical-thinking activities. *Journal of Teacher Education*, 56(1), 24–33. <https://doi.org/10.1177/0022487104272056>
- Weiss, R. E. (2003). Designing problems to promote higher-order thinking. *New Directions for Teaching & Learning*, 2003(95), 25–31. <https://doi.org/10.1002/tl.109>
- West, B. T., Berglund, P., & Heeringa, S. G. (2008). A closer examination of subpopulation analysis of complex-sample survey data. *Stata Journal*, 8(4), 520–531. <https://doi.org/10.1177/1536867X0800800404>
- Wilder, S. (2014). Effects of parental involvement on academic achievement: A meta-analysis. *Educational Review*, 66(3), 377–397. <https://doi.org/10.1080/00131911.2013.780009>
- World Economic Forum. (2015). *New vision for education: Unlocking the potential of technology*. http://www3.weforum.org/docs/WEFUSA_NewVisionforEducation_Report2015.pdf
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, 27(2), 172–223. <https://doi.org/10.1016/j.dr.2006.12.001>
- Zohar, A., Degani, A., & Vaaknin, E. (2001). Teachers' beliefs about low-achieving students and higher order thinking. *Teaching and Teacher Education*, 17(4), 469–485. [https://doi.org/10.1016/S0742-051X\(01\)00007-5](https://doi.org/10.1016/S0742-051X(01)00007-5)
- Zohar, A., & Dori, Y. J. (2003). Higher order thinking skills and low-achieving students: Are they mutually exclusive? *Journal of the Learning Sciences*, 12(2), 145–181. https://doi.org/10.1207/S15327809JLS1202_1

Author

HAJIME MITANI, is an assistant professor of educational leadership at Rowan University. His research interests include test score gaps, critical thinking skills, leadership skill requirements, leadership skill development, educator and leader preparation, education policy analysis, and program evaluation.