U.S. Middle School Mathematics Instruction, 2016

Heather C. Hill
Harvard University

Virginia S. Lovison
Harvard University

In recent decades, U.S. education leaders have advocated for more intellectually ambitious mathematics instruction in classrooms. Evidence about whether more ambitious mathematics instruction has filtered into contemporary classrooms, however, is largely anecdotal. To address this issue, we analyzed 93 lessons recorded by a national random sample of middle school mathematics teachers. We find that lesson quality varies, with the typical lesson containing some elements of mathematical reasoning and sense-making, but also teacher-directed instruction with limited student input. Lesson quality correlates with teachers' use of a textbook and with teachers' mathematical background. We consider these findings in light of efforts to transform U.S. mathematics instruction.

VERSION: August 2022

Suggested citation: Hill, Heather C., and Virginia S. Lovison. (2022). U.S. Middle School Mathematics Instruction, 2016. (EdWorkingPaper: 22-618). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/1h14-4f36

U.S. Middle School Mathematics Instruction, 2016

Heather C. Hill
Virginia S. Lovison
Harvard Graduate School of Education

The research reported here was supported by the National Science Foundation through grant #1417731 to Harvard University. The opinions expressed are those of the authors and do not represent views of the National Science Foundation. The authors would like to thank Dan McGinn, Dan Zahs, Barbara Gilbert, Carrie Maierhofer and Corinne Herlihy. The authors have no relevant financial or non-financial interests to disclose. Manuscript length: 9990 words.

Correspondence concerning this article should be addressed to Heather Hill, Harvard Graduate School of Education. Heather_hill@harvard.edu. Hill's ORCID: 0000-0001-5181-1573.

Abstract

In recent decades, U.S. education leaders have advocated for more intellectually ambitious mathematics instruction in classrooms. Evidence about whether more ambitious mathematics instruction has filtered into contemporary classrooms, however, is largely anecdotal. To address this issue, we analyzed 93 lessons recorded by a national random sample of middle school mathematics teachers. We find that lesson quality varies, with the typical lesson containing some elements of mathematical reasoning and sense-making, but also teacher-directed instruction with limited student input. Lesson quality correlates with teachers' use of a textbook and with teachers' mathematical background. We consider these findings in light of efforts to transform U.S. mathematics instruction.

Key words: instructional quality, teaching mathematics

Word count: 9,760

For most of the last 25 years, U.S. education leaders have advocated for more intellectually ambitious mathematics instruction in classrooms. This goal arose during the mid-1980s, as scholars developed new understandings of student learning (Bransford, Brown & Cocking, 2000; NCTM, 1989) and sought to re-orient classrooms away from teacher-driven procedural instruction to ones in which students participate in disciplinary practices, including mathematical sense-making and reasoning. By the early 2000s, these leaders had marshalled an array of resources to make this vision a reality: professional and political organizations wrote new standards for teacher and instructional quality including, in 2010, the Common Core State Standards (NCTM, 1989, 1991; National Commission of Science & Mathematics Teaching, 2000; National Mathematics Advisory Panel, 2008; National Governor's Association/Council of Chief State School Officers, 2010; National Research Council, 2001); state and local governments adopted or aligned to these standards, and provided teachers opportunities to learn about them; the federal government and private foundations supported the development of reform-aligned instructional materials; and professional developers sought to improve teacher knowledge and help teachers lead more rigorous mathematical inquiry in classrooms. Many anticipated that this community-wide work would improve instruction and student achievement in the subject.

Evidence about whether more intellectually ambitious mathematics instruction has filtered into contemporary classrooms, however, is largely anecdotal. Program evaluations have documented positive effects of specific efforts, for instance combining standards-aligned curriculum materials with professional development (e.g., Roschelle, et al., 2010) and providing teachers information on how students learn (e.g., Jacobs et al., 2007). Case studies and small-sample quantitative studies found variability in instructional quality, but also evidence of high-quality lessons (e.g., Lloyd & Wilson, 1998; Hill et al., 2008; O'Connor et al., 2015). Yet other evidence suggests that mathematics classrooms, on average, still fall short of reformers' goals. An evaluation of standards-based mathematics professional development found that teachers typically elicited substantive student mathematical thinking just two to three times per hour of instruction (Garet, et al., 2011). The Measures of Effective Teaching (MET) study (Kane & Staiger, 2012) depicted mathematics instruction in six urban districts as largely absent mathematical sense-making, mathematical practices and student cognitive challenge, and Hill, Litke & Lynch (2019) paint a similar picture in two of five urban districts.

Despite this large volume of research in recent decades, few comprehensive portraits of teacher and teaching quality in U.S. mathematics classrooms exist. The studies referenced above, as well as other case studies appearing in journals and monographs, are typically conducted with either small or non-representative samples in a handful of districts, limiting the generalizability of this evidence. The TIMSS study, which did feature a nationally representative sample (Hiebert et al., 2005) collected its data in 1999, prior to the Common Core and before significant efforts in the 2000s to provide teachers with new curriculum and professional development.

We argue that more recent information on national progress toward more ambitious instruction in mathematics classrooms would be useful. Because efforts to transform mathematics classrooms have persisted since the late 1980s, they constitute one of the longest-running efforts to substantially alter U.S. instruction. The TIMSS study (Hiebert, et al., 2005) suggested that instruction as captured in 1999 did not yet meet reformers' goals. Providing an updated look at mathematics classrooms may help provide evidence about whether the passage of time – and the presence of ongoing efforts to transform instruction – has resulted in a more optimistic picture of the rigor of classroom mathematics instruction.

To this end, this study collected videotaped lessons from a random sample of U.S. middle school mathematics teachers, and scored those lessons using a standardized observational instrument. We describe the background, procedures and results for our study below.

Background

Prior to the standards-based mathematics reforms of the 1990s, mathematics instruction was reported to be uniformly drab. For instance, writing about the instruction early reformers hoped to change, Cohen and Ball (1990) observed:

> Mathematics teaching in most elementary classrooms emphasizes rules, procedures, memorization, and right answers (Goodlad, 1984; Stodolsky, 1988). Students seldom confront serious mathematical problems and are rarely expected to reason about mathematical ideas. Teachers stand at the board, show students how to do a particular procedure or type of problem, and assign practice exercises. Students then work quietly on these, asking the teacher for help if they get stuck. When students are done, the teacher checks their answers, marks the ones that are wrong, sometimes goes over the steps once again, and then students fix their incorrect answers. (Cohen & Ball, 1990, p. 234)

This view was supported by findings from teacher surveys from the time, which suggested that U.S. mathematics classrooms featured superficial, repetitive treatment of content (Porter, 1989).

Beginning in the mid-1980s, reformers proposed newly ambitious goals for mathematics teaching and learning (California State Department of Education, 1985; NCTM, 1989; 1991; 2000; NGA/CCSSO, 2010).). Many new standards contained two important dimensions. First, they asked teachers to portray the subject in a more disciplinary light, focusing on mathematical practices (e.g., justification) and representing the subject as one that requires sense-making and understanding. Second, new standards asked teachers to actively involve students in these practices and sense-making. Instead of receiving knowledge, students were to construct it through reasoning and application; instead of presenting rules and formulas, teachers were to set the stage for student growth through problem-posing and encouraging mathematical discussion and debate. Following Cohen (2012), we refer to the ideas in these standards and in the policy-making and mathematics education community more generally as “ambitious instruction.”

Early studies of standards-based reform suggested that using policy to induce such changes would not be an uncomplicated process. Studies of California’s initial standards-based reforms (Ball, 1990; Cohen, 1990; Wilson, 1990) indicated that teachers interpreted novel policies as compatible with more traditional curricula and pedagogy, and enacted a mix of new and old in their classrooms. Later analyses (Spillane & Zeuli, 1999) suggested that teachers added standards-based reform to the margins of practice while leaving the core of their instruction unchanged. And by the end of the decade, data collected for the TIMSS video study (Hiebert et al., 2005) showed that these reforms had not entered most U.S. middle school classrooms in meaningful ways. TIMSS authors observed lessons characterized by low levels of applied problems, the use of simple procedures for solving problems, and little sense-making and disciplinary justification. Another study completed around the same time period (Weiss et al., 2003) found that many middle school lessons lacked mathematical rigor, student intellectual engagement, and student sense-making. Some featured inaccuracies in content, and observers sometimes noted that teachers appeared to lack understanding of the content.

Scholars quickly focused on an explanation for relatively static pedagogy: that teachers lacked the resources necessary to carry off ambitious instruction. One key resource was teachers' content knowledge—sometimes known as "mathematical knowledge for teaching" (MKT) (Ball, Thames & Phelps, 2008; Thompson & Thompson, 1996). Scholars documented how low levels of such knowledge constrained teachers' instructional decision-making, led to missed opportunities for sense-making and explanations, and failed to spark student reasoning and inquiry (e.g., Cohen, 1990; Heaton, 1992). The second instructional resource was curriculum materials (Ball & Cohen, 1996; Davis & Krajcik, 2005). Prior to 1990, most textbooks contained guidelines for teacher explication of content and exercises for student practice.

As scholars consensed on the centrality of these two resources in supporting ambitious instruction, practitioners and policy-makers set out to provide both to teachers at scale. Teachers' mathematical knowledge was the intended target of NCLB's teacher quality regulations, which required middle school mathematics teachers to obtain a subject matter major, equivalent credits to that major, or to pass an exam demonstrating competency in the subject. However, a companion study to this one (Hill et al., 2019) found that the fraction of middle school teachers with a mathematics degree decreased between 2005 and 2016. The development of curriculum materials meant to support ambitious instruction was funded in the 1990s by the National Science Foundation (NSF) and, following the introduction of the Common Core in 2010, by private foundations. These materials reflect students' natural ways of thinking about mathematics, promote an inquiry approach to new topics, and are designed to enhance student facility in the application of mathematical concepts and ideas (Stein, Remillard & Smith, 2007). However, the companion study found only a fifth of U.S. middle school teachers used such materials, and that most teachers reported mixing the use of published materials with teacher-created materials, including those made themselves or with colleagues and those found on internet sites like teacherspayteachers.com.

Recent evidence about whether classrooms have shifted to offer students more access to ambitioius instruction is also similarly mixed. An observational study found almost uniformly traditional instruction among participating mathematics teachers (Kane & Staiger, 2012). However, a smaller study of five urban districts shows that in two, standards-aligned instruction did exist in elementary classrooms (Hill, Litke & Lynch, 2019). Further, both studies also revealed the presence of teacher mathematical errors within the classroom – some consisting of minor misuse of mathematical terms, and others more serious, like incorrectly solving problems or mis-stating mathematical ideas and concepts. These errors are consistent with reformers' concerns about teachers' weak subject matter knowledge.

Neither of these observational studies, however, included representative samples, meaning that these findings cannot be generalized beyond the specific districts participating in the studies. To gather information from such a sample, this study contacted middle school mathematics classrooms from across the country and recorded instruction in 98 of those classrooms in 2016. From this data as well as a brief post-observation survey, we can ask four questions. The first two gauge instruction vis-à-vis reformers' goals and concerns:

1. To what degree do observed lessons feature ambitious instruction?
2. What is the frequency and severity of teacher mathematical errors in the classroom?

A third question is meant to provide a rough comparison between lessons collected toward the beginning of the reforms and our dataset, collected several years after the Common Core:

3. How does mathematics instruction in 2016 compare to typical instruction in 1999?

Finally, data on the post-observation survey allowed us to ask:

4. What instructional resources characterize 2016 lessons with higher instructional quality?

This study is descriptive, as are many other studies of its kind (Hiebert et al., 2005). Yet without strong description, we cannot successfully design new policies and supports for improvements in teaching and learning.

Methods

Sample

We aimed to achieve a nationally representative sample of middle school mathematics teachers and the lessons they teach. We defined middle schools as public schools containing at least 10 students in at least one of the grades 6, 7 and 8. Using the NCES Common Core of Data, we identified 24,270 schools that met these criteria, stratified those schools by region and size, then selected schools at random, where the probability of selection was proportional to school size within region. We then sampled 1,822 schools from this larger set, completed a district IRB where necessary, and gathered teacher names and contact information from school websites and via phone calls. We obtained teacher rosters from 1,583 schools, or 87% of the original sample. Within these schools, we selected a single teacher at random to participate.

We conducted an initial “light touch” study with this sample in 2014-2015 using the Dillman Total Survey Response method of repeated contacts via U.S. mail and email, and a stipend of $300 to compensate teachers for recording four lessons with video equipment mailed to them by us. However, this study yielded a very low teacher response rate (11%). In 2016, we abandoned this initial sample, selected 200 new mathematics teachers from districts that had previously approved the “light touch” study, and began a “high touch” recruitment process, with letters, personalized phone calls, and a $300 stipend for collecting a single classroom video. When response rates remained low, we expanded this sample to include a total of 442 teachers from the approved districts. Ultimately, 109 of those teachers agreed to participate in the study. We received usable video from 98 of these teachers, for a teacher-level response rate of 22%. Five of those videos could not be scored due to poor audio or video quality, leaving 93 in our analytic pool. For more details on the sampling and recruitment methods, please see (Zahs et al., 2018).

This response rate was well below what we were hoping to achieve, given the resources devoted to reaching and recruiting teachers. In speaking with declining teachers by phone, teachers consistently cited several reasons for their refusal: a lack of time, changing mathematics standards and curricula, worry about their ability to meet those standards, and concerns that the video would be used for evaluative purposes. Our study also required teachers to deploy a remote camera unit, which may have felt more burdensome than allowing a videographer – as previous video studies such as TIMSS asked – to record teaching. Overall, we succeeded in converting very few refusals into consents. The challenges we faced highlight the difficulty of recruiting teachers in today's school environments and are consistent with the experiences of researchers in other fields who have also faced declines in response rates (Meyer, Mok & Sullivan, 2015).

Information about our 2016 achieved sample (in Table 1), and the extent to which it departs from a companion nationally representative survey that achieved a 58% response rate around the same time (Hill, Kelly-Kemple & Lovison, 2019) informs the degree of risk in interpreting our results. On average, teachers in both the video and survey studies had between 10 and 11 years of experience. In both samples, a majority of teachers held undergraduate and/or graduate degrees in education, and the samples were similar in the proportion of teachers who held a graduate degree in mathematics. However, teachers in the video sample are slightly less likely to have an undergraduate mathematics major/minor, suggesting that comfort with mathematics was not a strong factor driving selection into the study. However, we cannot rule out other factors that might drive selection, for instance perceived instructional expertise or the use of specific curriculum materials.

Table 1
Characteristics of teachers, 2016 video sample vs. 2016 national survey

	2016 video sample mean	National survey sample mean
Years of experience	10.62	11.06
	(6.79)	(7.84)
Math major	0.19	0.27
	(0.40)	(0.45)
Math minor	0.13	0.17
	(0.34)	(0.38)
Graduate degree in mathematics	0.09	0.07
	(0.28)	(0.25)
Education major	0.58	0.64
	(0.50)	(0.48)
Graduate degree (education)	0.60	0.59
	(0.49)	(0.49)
Observations	93	904

Note: Standard deviation reported in parentheses.

In addition to collecting novel data for this study, we also obtained all the original 1999 TIMSS videos (83 in total) from the National Center for Educational Statistics (NCES). We then rescored each video using our own observational instrument so that we would have a common metric with which to describe differences in instructional practices between the two samples. Information on the TIMSS sampling procedures can be found in Jacobs and colleagues (2003). However, because the response rate for the current study is well below that of the TIMSS video study (93%), our ability to draw comparisons across the two samples is limited. We thus urge caution in the interpretation of results from this study and discuss our findings in light of differences in response rates.

Lesson capture

We asked teachers to collect video from the first class scheduled after 10 AM on the Tuesday after they received our video capture equipment. Teachers recorded their own video, placing a SWIVL video unit with iPad and embedded microphones toward the front of their classroom to capture boardwork. Teachers also wore a microphone on a lanyard. Most teacher and student speech was audible enough for coding purposes, and most videos lasted between 45 and 60 minutes.

Post-observation Survey

In both 1999 and 2016, teachers responded to a brief survey following lesson capture. These surveys asked teachers to report on the origin of the materials used in the lesson, and to indicate whether they majored or minored in mathematics as an undergraduate, or obtained a degree in mathematics as a graduate student. In the 1999 sample, eight teachers failed to return surveys. In 2016, all teachers with usable videos also returned surveys.

Video Scoring

We scored both the 2016 and 1999 TIMSS videos using the *Mathematical Quality of Instruction* (MQI) instrument. This instrument measures many of the elements of ambitious instruction, including the presence of disciplinary features (e.g., mathematical explanation, generalization, language precision), the role students play in the production of knowledge (e.g., student explanation, cognitive demand of student tasks) and classroom discourse (e.g., teacher uptake of student ideas). The MQI also has well-established technical properties and a demonstrated relationship to teacher knowledge and student outcomes (see Hill, Kapitula, & Umland 2011; Hill, Umland, Litke & Kapitula, 2012; and Kelcey, Hill & Chin, 2017 for a description of instrument validity and score validation efforts). We randomly assigned two trained raters to watch and score each lesson, with lessons broken into 7.5-minute segments on each of 16 MQI items (see Table 2 for items and brief definitions) using a *Not Present* (1) to *High* (4) scale. Although the meaning of this scale varied slightly by item, *Not Present* (1) generally indicated the activity described by the item did not occur; *Low* (2) generally indicated a brief or superficial instance of the activity; *Mid* (3) generally indicated a lengthier or more substantive implementation of the activity; and *High* (4) generally indicated lengthy and strong implementation of the activity. To illustrate how this rubric looked when mapped to specific items, Table 2 presents the *Low* (2) and *High* (4) score points for each item.

Table 2

Dimensions and items in the Mathematical Quality of Instruction (MQI) instrument

Name of dimension/item	**Definition**	**Low (2)**	**High (4)**
Richness: **Presence of mathematical meaning (e.g., sense-making, explanations) or disciplinary practices, regardless of whether they are enacted by the teacher or student.**			
Linking Between Representations	Teachers and/or students link and connect different representations of a mathematical idea or procedure.	Links are present in a pro forma way.	Links and connections are explicit, with extended, careful work characterized by elaboration and detail.
Explanations	Teacher and/or students focus on why a procedure works, why a mathematical statement is true, etc.	A mathematical explanation occurs as an isolated instance in the segment.	One or more mathematical explanation(s) is a focus of instruction in the segment.
Mathematical Sense-Making	Teacher and/or students pay attention to the meaning of quantities (e.g., 7/8ths) definitions and procedures, whether solution methods and answers make sense, etc.	Teacher and/or students focus briefly on meaning.	Teacher and/or students focus on meaning in sustained way during segment.
Multiple Procedures or Solution Methods	Teacher and/or students use multiple solution methods to solve a single problem (including shortcuts), or multiple procedures for a given problem type.	Teacher or student briefly mentions a second procedure or method, but the method is not discussed at length or enacted.	Explicit comparison of multiple methods for efficiency, appropriateness, ease of use.
Patterns and Generalizations	Teacher and/or students *first* examine examples, *then* use information from examples to develop a mathematical generalization, property, or definition.	There is brief work on developing a generalization or building a definition, but this work is undeveloped and/or is not the primary focus of the segment.	The pattern or generalization is codified, AND the work is complete, clear and detailed.

Mathematical Language	Teacher fluency in using mathematical language; whether the teacher supports students' use of mathematical language.	Low density of mathematical language OR moderate density but sloppy use.	High density of mathematical terminology, including explicitness around the meaning of terms, pressing students to use mathematical language.
Working with students: **Teachers' use of student mathematical ideas and misunderstandings during the segment.**			
Remediation of Student Errors	Teacher addresses student misunderstandings.	Procedural remediation; brief conceptual remediation.	Teacher engages in conceptual remediation *systematically* and *at length*.
Teacher uses Student Mathematical Contributions	Teacher uses student contributions to develop the mathematics of the lesson.	Teacher responds in a pro forma way to student contributions.	Students' mathematical ideas are woven at length into the development of mathematical ideas during the segment.
Errors**: Teacher errors in the presentation of content.**			
Mathematical Content Errors	Teacher solves a problem incorrectly; defines terms incorrectly.	A brief content error. Does not obscure the mathematics of the segment.	Content errors occur in most or all of the segment.
Imprecision in Mathematical Language or Notation	Teacher uses mathematical terms or notation imprecisely.	Brief instance of imprecision. Does not obscure the mathematics of the segment.	Imprecision occurs in most or all of the segment.
Lack of Clarity in the Mathematics	Teachers' presentation of a mathematical point is muddled, confusing, or distorted; errors make it difficult to discern the point of the segment.	Brief lack of clarity. Does not obscure the mathematics of the segment.	Lack of clarity occurs in most or all of the segment.
Common core aligned student practices: **Extent to which students participate in the development of the mathematics by asking questions, providing explanations, engaging with cognitively demanding tasks.**			
Students Provide Explanations	Students explain why a procedure works; explain what an answer means; explain why a solution method is suitable or better than another method.	One or two brief student explanations are present.	Student explanations characterize much of the segment.
Student Mathematical Questioning and Reasoning	Students make conjectures, counter-claims, ask mathematically motivated questions, engage in reasoning, etc.	One or two instances of brief student mathematical questioning or reasoning are present.	Student mathematical questioning or reasoning characterizes much of the segment.

Students Communicate about Mathematics	Extent to which students communicate their mathematical ideas, either in whole-group or small group settings.	Student contributions are very brief, e.g. one- or two- word answers, but occur regularly during the segment.	Substantive student contributions characterize the segment (e.g., students present solution methods, discuss mathematical ideas).
Task Cognitive Demand	Whether tasks, as enacted, require students to think and reason about mathematics.	There is a brief example of a cognitively demanding activity (e.g., think-pair-share to define a term).	Students engage at a *high* level of cognitive demand (e.g., develop a generalization, provide explanations, link representations).
Student Work with Contextualized Problems	Students work with story problems, real-world applications, experiments that generate data.	The contextualized problems are executed as mostly rote/routine exercises.	Students are allowed significant opportunities to think and reason about contextualized problems.

Note: For all codes, “*Not Present*” is scored as a 1. A score of *Mid* (3) indicates a lengthier or more substantive implementation of the activity occurred during the segment, but did not meet the criteria for *High* (4).

Raters scored every 7.5-minute segment in each lesson, including final segments that were more than a minute long. We also asked raters to assign an *Overall MQI* score to each lesson, where scores of 1 and 2 indicated mathematically problematic lessons, a score of 3 indicated either a purely procedural lesson or a lesson that had a mix of positive (*Richness*, *Common Core Student Practices*) and negative (*Teacher Errors*) elements, and scores of 4 or 5 indicated more meaning-oriented mathematics featuring significant student participation. These scores correspond to variability in existing U.S. instruction, because the MQI instrument was developed to capture existing variability rather than judge instruction against an ideal.

To help control for rater effects, we used the same six raters to score both datasets, and kept the percentage of lessons scored per rater in rough proportion across both samples. However, we could not blind raters to the origin of the lesson because of differences in video capture procedures and scoring platform. Because the MQI instrument has strict scoring guides, however, we expect that halo effects associated with data source would be minimized. Raters agreed on *Overall MQI* scores for 54% of lessons (kappa =.24) and 73% of segment-level item scores (kappa = .47). Raters' scores differed by no more than one point for 98% of *Overall MQI* scores and 96% of segment-level item scores.

Analyses

Because we collected only one lesson per teacher, we cannot characterize teachers' instruction; instead, we focus on characterizing the lessons included in the sample. To characterize the 2016 lessons, we first present *Overall MQI* ratings. Next, we make use of segment-level data in two ways. First we present the raw percentages at the item-by-segment level. This answers the question: within a given segment, what is the likelihood of the activity captured by a specific MQI item appearing at a specific level of quality? Next we calculate the highest score achieved within a lesson on each item. This answers the question: did the activity captured by the MQI item appear at some point during the lesson, and if so, at what level?

The fact that two raters viewed each segment and lesson complicates our reporting here, introducing half-scores (e.g., a score of 4.5 on *Overall MQI*) and yielding lumpy and unsightly distributions. We took two pathways to ameliorate this issue. For the *Overall MQI* and when aggregating segment data to the lesson level, we present the lesson's highest score from either rater. This "best rating" approach recognizes that lesson quality is measured with error introduced by rater perceptions and biases, and credits classrooms where higher-quality instruction may have occurred. For the raw segment-level data, we randomly sample one rater per segment for the purposes of presentation. Comparisons of means and standard deviations in the full datasets vs. the data presented above suggests very few substantive differences.

All bivariate comparisons were conducted using data from both raters. To answer our third research question and estimate differences between the 1999 and 2016 samples, we use the following mixed effects model:

$$Score_{ijk} = \beta_0 + \beta_1 Year_k + u_k + \eta_i + \epsilon_{ijk}$$

where $Score_{ijk}$ is the score assigned for an item by rater *i* for segment *j* in lesson *k,* YEAR is a binary variable equal to one if the lesson was drawn from the 2016 sample and zero otherwise, u_k is a lesson random effect and η_i is a rater random effect. We assume both the lesson and rater random effects are normally distributed. The coefficient of interest is β_1, which captures the average difference in item scores between the 2016 and 1999 samples. We conduct two different kinds of analysis using this model. We begin by averaging MQI items into two factor composites at the rater-segment level: *Ambitious Instruction* (*Richness*, *Working with Students*, and *Common Core-Aligned Student Practices [CCASP])* and *Errors*. These composites are indicated by other analyses of the MQI data (Blazar, et al., 2017) and are more sensitive to specific dimensions of instruction than *Overall MQI.* For these aggregate analyses, $Score_{ijk}$ is the average over all the relevant item scores in segment j of lesson k. Then, to help illuminate patterns underlying these results, we run separate regressions for each item on the MQI rubric. In this analysis, $Score_{ijk}$ is a particular item.

We answer our fourth research question by reporting whether *Overall MQI* scores differed by whether the teacher reported using a textbook, and whether the teacher reported a mathematics degree.

Finally, to give readers a better descriptive sense of lessons at each level of *Overall MQI*, we reviewed about 15% of scored lessons and then selected three to briefly present below: one that represents the lower end of MQI quality (a score of 1 from one rater, 2 from another), one that represents the middle portion of MQI quality (a score of 3 from both raters), and one that represents the higher end of MQI quality (a score of 4 from both raters).

Note we could not apply weights to the MTTS sample due to the transition between the light touch (where teacher probabilities of entry to the sample could be calculated) and high touch (where probabilities of entry to the sample could not be calculated) samples. For consistency, we did not apply TIMSS weights.

Results

RQ1 & RQ2: To what degree do observed lessons feature ambitious instruction? Teacher mathematical errors?

MQI Ratings

Figure 1 represents the distribution of the 2016 lessons' highest score from either rater on *Overall MQI.* We found that 15 lessons, or 16%, were mathematically problematic enough to warrant an MQI score of 1 or 2. Twenty-eight lessons (30%) featured enough disciplinary practices and student participation in those practices to warrant an *Overall MQI* score of 4, and one lesson scored a 5 (1%). And, 49 lessons (53%) were more traditional in nature or a mix of positive (*Richness*, *CCASP*) and negative elements (*Errors*).

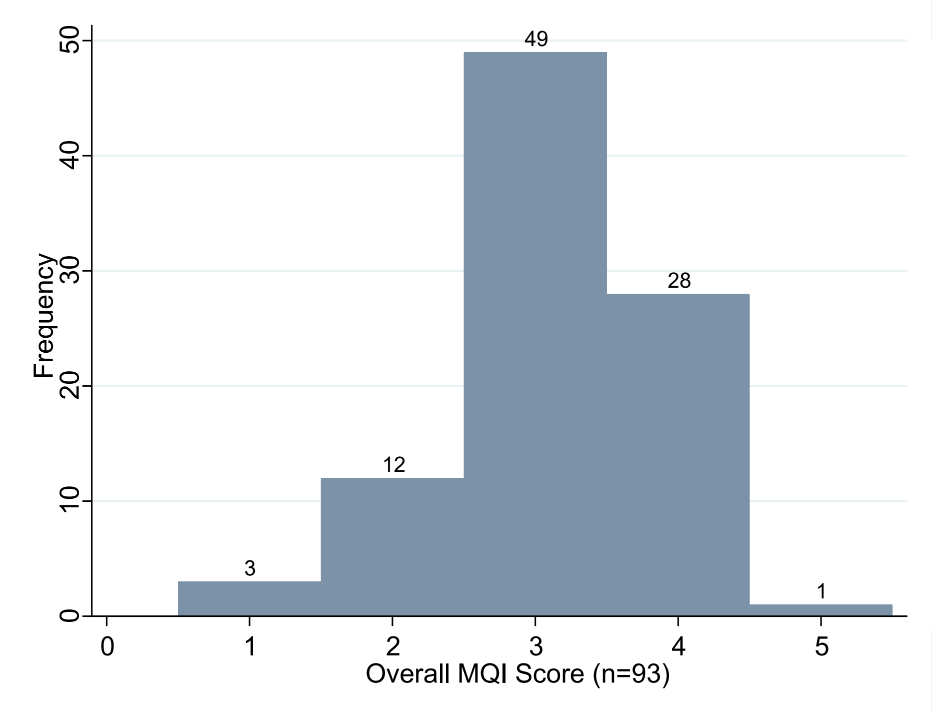


Figure 1
Overall MQI Score

Tables 3A and 3B describe the data in more detail. We use *Linking Between Representations* as an example in how to interpret these tables. Table 3A shows that linking between representations is relatively rare at the segment level – only 15% of segments feature this practice. However, a majority of lessons (58%) have at least one example of this practice. Both tables show that when this practice appears, it generally does so at a low level.

The remaining practices included in *Richness of the Mathematics* follow this pattern. Most lessons saw each of several major practices – linking, explanations, sense-making, multiple solution methods – at least once (Table 3B). However, the overall prevalence of these activities was not high across segments, and when present, they tended to score a *low*, indicating they were brief or superficial (Table 3A). Scores of *high*, which indicate extended and strong implementation of the practice, were particularly rare. The same was true for many *CCASP* items. Though student explanations, student reasoning, and working on cognitively demanding activities and contextualized problems occurred at some point during most lessons (Table 3B), these occurrences tended to be brief (Table 3A) or, in the case of contextualized problems, not cognitively demanding. For *Students Communicate About Mathematics*, scores of *low* represents traditional practice, such as teacher-led instructional discourse.

In *Working With Students*, both the segment and lesson level analyses demonstrate a fair amount of *Remediation of Student Errors*, though that remediation tended toward the procedural (*low*) rather than conceptual (*mid* or *high*). Similarly, *Teachers Use Student Contributions* typically occurs in a pro forma way – evaluating student responses and moving on (*low*), and less often weaving students' mathematical ideas into the fabric of the lesson (*mid* or *high*) (Table 3A). While most segments were free of *Content Errors*, about 10% of lesson segments contained at least some *Imprecisions in Language*, and about 10% of segments saw a *Lack of Clarity (*Table 3A*)*. Analysis at the lesson level, however, shows that about 42% of lessons contained a mathematical content error, and about half contained a moment that lacked clarity or a problem with mathematical language.

Table 3A
Percentage of Segments at Each Score Level on MQI Items

Item	Not Present	Low	Mid	High
<u>Richness</u>				
Linking Between Representations	85	10	5	1

Explanations	92	6	1	1
Math Sense-Making	54	24	20	3
Multiple Solution Methods	89	7	4	1
Patterns and Generalizations	97	1	1	1
Math Language	6	31	50	13
Working with students				
Remediation of Student Errors	55	34	10	1
Teachers Uses Student Contributions	25	56	17	1
Errors				
Math Content Errors	93	3	3	<1
Imprecision in Language	89	8	3	<1
Lack of Clarity	89	6	4	<1
Common-core aligned student practices (CCASP)				
Student Explanations	87	10	3	0
Student Reasoning	90	10	1	<1
Student Communication	19	46	33	2
Task Cognitive Demand	61	28	10	1
Contextualized Problems	66	24	7	2
Number of segments	557			

Note: Cases where numbers do not sum to 100 are due to rounding.

Table 3B

Percentage of lessons with best rating, lesson-wide, of a NP, Low, Mid, or High

Item	Not Present	Low	Mid	High
Richness				
Linking Between Representations	42	27	23	9
Explanations	52	37	6	5
Math Sense-Making	4	18	51	27
Multiple Solution Methods	47	22	24	8
Patterns and Generalizations	84	6	4	5
Math Language	0	5	39	56
Working with students				
Remediation of Student Errors	6	43	44	6
Teachers Uses Student Contributions	3	29	60	8
Errors				
Math Content Errors	58	24	15	3
Imprecision in Language	45	37	17	1
Lack of Clarity	53	25	19	3
Common-core aligned student practices (CCASP)				
Student Explanations	33	51	16	0
Student Reasoning	42	47	10	1
Student Communication	3	8	82	8
Task Cognitive Demand	12	51	34	3
Contextualized Problems	39	32	20	9

Number of observations	93

Lesson Snapshots

To provide additional context for the scores presented above, we describe lessons at *Overall MQI* scores of 1.5, 3 and 4.

Sample Lesson 1 (MQI Score 1.5). This lesson featured teacher and students working on solving single-variable equations, with some meaning-oriented mathematical practices along the way. However, not all of the information conveyed in the class was mathematically accurate. The teacher began by asking students a terminology question:

> T: First, how do you recognize that this example ($x -2 = 3$) is an equation?
>
> S: There's an equals sign.
>
> T: There's an equals sign.

Using mathematical language precisely is a key curricular standard and mathematical practice, and the teacher reinforced this at the start of this lesson, first here, and then by asking students to similarly define other terms (e.g., inverse operations).

The teacher next asked students to solve $x - 2 = 3$. She walked students through the calculation, strongly scaffolding student responses:

> T: $x - 2 = 3$. If you had to do this…what would be the first question you would ask yourself?
>
> S: Try to find what x is.
>
> T: What x is. In order to find what x is, we need to find the inverse operation. What is the inverse operation of subtracting 2?
>
> S: Adding two?
>
> T: Adding two. So every student go ahead and add two to both sides.

In the above excerpt, the teacher prompted students to use the inverse operation, and then pointed to exactly how: by doing the opposite of subtracting two. Then the teacher interjected a moment of meaning into the conversation:

> T: Adding two. Let's add two to both sides. *The reason is that we want to make sure both sides are balanced.* So when I add, I've isolated my variable, which is x.

By pointing out the need to balance equations, the teacher briefly referenced a key tenet of algebra problem solving – keeping the value of both sides of an equation the same. The teacher and students then completed solving the problem and checked their answer by substituting five for x. Thus in these opening moments of the lesson, there was a brief drive for language precision and brief sense-making around how completing the same action on both sides of an equation leaves that equation balanced.

A few moments later, however, a mathematical misconception entered the lesson. As students worked on $9 = a - 5$, the teacher asked them to rewrite the equation as $a – 5 = 9$, and then justified that action by referring to the commutative property:

> T: Here's my question. We get this type of equation ($9 = a - 5$) all the time. How can I rewrite this equation? And tell me how you know. How can we rewrite it in a way that will make it easier to solve?
>
> S: $a – 5 = 9$
>
> T: $a – 5 = 9$. Thumbs up if you can agree that we can write it like this…[students give thumbs up]. Why? Why can we rewrite any equation that looks like this? Looking for an academic word here. What am I looking for?
>
> S: The commutative property.
>
> T: The commutative property. What does the commutative property tell me I can do to an equation? …What does the commutative property state that I can do to equations that look like this to help make life easier when you solve it?... I can use the commutative property to rewrite it because. Because what does the commutative property say I can do?

> S: [reading from notes] The order in which two numbers are added or multiplied does not change the sum or product.
>
> T: So I can rewrite it, right, to put the variable *a* in front, so that when I solve this, I can actually solve it and get my answer.

Although justifying procedural steps using mathematical properties is both fundamental to mathematics and can help students make sense of those steps, this teacher's attempt to do so muddied the waters. To start, the procedure the teacher executed is properly justified by the symmetric property of equality (if $a = b$, then $b = a$), not the commutative property. Second, this procedure adds an unecessary step to the problem's solution and may reinforce rigid notions about the equals sign held by some students. Yet the teacher repeated the same assertion about commutativity to justify rewriting equations throughout the remainder of the lesson.

As with other lessons scoring a 1 or 2 on *Overall MQI*, the teacher did most of the talking, with students offering one-word or short phrases as answers to direct questions. Student tasks were generally of low cognitive demand, for instance recalling information, reading from notes, or hazarding guesses to the teacher's questions.

Sample Lesson 2 – MQI Score 3. This lesson featured brief moments of mathematical meaning and very occasional questions that challenged students to think, but also generous amounts of relatively procedural, teacher-centered instruction. Teachers and students began the lesson by working on graphing the inequality $x \geq -2$ on a number line. The teacher asked what this statement meant to students, specifically, what numbers would satisfy the inequality. Students replied with several values: $1, -2, -1$, and the class briefly reviewed how to draw and graph each on a number line. The teacher and students then considered how to graph this particular equation, locating 0 and -2 then asking:

> T: Now we're going to graph this (equation). Remember you have that circle (on *–2*). Is that circle going to be open or is it going to be closed? What determines whether a circle is open or closed?
>
> S: Negative or positive.
>
> T: Not if it's negative or positive.
>
> S: Equal to or?
>
> T: Well, this little guy right here [teacher circles the long dash that is part of the inequality symbol]. If it is equal to, it's going to be colored in, it's going to be a solid dot. Then you have to decide which way to go. …. In this case, am I going right or left?
>
> S: Right.
>
> T: [draws the line as directed by the student]

Though the lesson began with brief sense-making about the meaning of $x \geq -2$, the teacher provided no explanation that a solid endpoint is used to include the number *2*, or a connection of the finished product back to the sense-making students had done earlier. She did, however, continue to make sense of inequalities several times as the lesson went on, for instance saying for $x > 24$, "…this is telling me, a number greater than 24" and while graphing $y > -3/5x + 6$, helping students be strategic about the plotting of points:

> T: Now, this equation says that *y* is gonna be greater than anything that's on this side. But the problem with that is we don't know what side of the line that means. It means that we can have solutions that are all on this side of the line, or I could have solutions that are all on this side of the line….So you have to kind of crank the numbers at this point….In order to do that, we're going to have to substitute a single point on my whole coordinate grid. You can choose any point that you want to, but there are always going to be some points that are easier and some points that are harder. Because if I choose points that are down here in my third quadrant, then I end up having to deal with a lot of negatives, right?

The lesson also contained brief moments of student cognitive demand – mainly in the form of teacher questions to students that included wait time for students to think and respond.

However, a majority of the lesson followed an "inquiry-response-evaluation" format, with the teacher calling for steps in the graphing procedure and students supplying them:

> T: Now once we hit 8th grade, we're starting with inequalities that have two variables. We have usually an x and a y.
>
> Student: Whoa.
>
> [Teacher writes $3x + 5y > 30$]
>
> T: ….It's the same thing. Same thing that we've been working on. If we pretend [the inequality symbol] is just an equals sign, how would we solve this? What would our equations have to look like to graph it?
>
> S: $y = mx + b$.
>
> T: So what form?
>
> S: Slope-intercept.
>
> T: You gotta know this. So subtract the $3x$ on both sides….

The teacher and students continued to isolate y, then students graphed the equation on whiteboards while the teacher circulated checking their work. Throughout, student contributions tended to be brief, fill-in-the-blank answers, and the teacher's use of those contributions focused on evaluating those answers and moving students onto the next step.

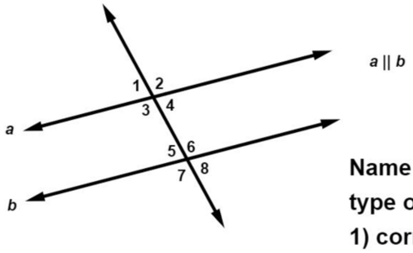
Other lessons scored as a 3 featured the same pattern described here – occasional moments of richness or student mathematical contributions, but little sustained focus on mathematical ideas, or the weaving of students' own mathematical ideas into the class discussion. Still other lessons with an MQI score of 3 simply didn't cover much content. In one, students competed to solve math problems and thus earn the right to make a free throw in the classroom basketball hoop. In another, a teacher used an lesson found in a *National Council of*

Teachers of Mathematics journal, *Barbie Bungee Jumping*, but focused mainly on data collection and graphing results rather than making predictions or interpreting data (see Litke & Hill, 2020).

Sample Lesson 3– MQI Score 4. This lesson featured substantial sense-making around geometry, including informal arguments about the angles created when parallel lines are cut by a transversal. Often, students provided these arguments, and the teacher elaborated on students' responses. The class was working with the following figure:

Figure 2

Angle Measure Activity



Name a pair of each type of angle listed.
1) corresponding
2) vertical
3) alternate interior

After a brief review of the mathematical terminology in the task, the teacher invited students to reason about the pictured angles:

> T: What if I told you that this angle right here, Angle 2 measures 110 degrees. If I told you that, what else could you tell me?
>
> S: That Angle 1 measures 70 degrees.
>
> T: Good. Angle 1 measures 70 degrees. But it's not enough to just say that. Right?
>
> S: Because it adds up to 180 and that's supplementary.
>
> T: Because Angle 1 and Angle 2 are supplementary angles, it must add up to 180. Good. What else do we know? Just because I gave you that one teeny tiny measurement, what else do we know?

Here, an intercom announcement leads to members of the track team exiting the room. In this short excerpt, however, we can already see the teacher pressing students for explanations – why Angle 1 measures 70 degrees – and also revoicing student talk with more precision ("Because Angle 1 and Angle 2 are supplementary…."). Instruction picks up again with a response to the teacher's query:

> S: That 3 and 4 have to be 180.
>
> T: That 3 and 4 have to equal 180 too. They're gonna be supplementary also. Do I know anything about the measurements of 3 and 4?
>
> S: Angle 3 has to be 110.
>
> T: Angle 3 has to be 110 degrees.You're absolutely right. Why?
>
> S: Because it's across from Angle 2.
>
> T: Stop. It's across from angle 2. And they are both what?
>
> S: And they're both situated on the same line.
>
> T: OK, when they are across from each other, they are what kind of angle pair?
>
> S: Vertical.
>
> T: They are vertical. When lines are parallel, we know that angles that are vertical are congruent. Actually, you don't even need to have parallel lines to have vertical angles be congruent. Vertical angles will always be congruent.

In this excerpt, we see many features of high-quality instruction: several student explanations, dense student use of mathematical vocabulary (vertical and supplementary angles, congruent) and some depth – even in this warm-up activity – in the treatment of the topic. These features continued throughout the lesson.

Reviewing the lesson as a whole suggests, however, that these standards-aligned features occurred within a more conventional instructional format. The first half of the lesson involved going over warm-up and homework problems; the second half featured the teacher handing out a worksheet to give students "some time to work." Discourse generally pingponged between the teacher and a single student at a time and focused on correct answers and reasoning, and the cognitive demand of the problems students solved did not appear high. These features were common to other lessons scoring above average on *Overall MQI*, and were one reason that only one lesson scored a five.

RQ3: How does mathematics instruction in 2016 compare to typical instruction in 1999?

Tables 4 and 5 compare MQI item ratings for the 2016 and 1999 samples. Table 4 shows that we observed more ambitious instruction in the 2016 sample than in the 1999 sample (Table 4). On average, 2016 lessons scored 0.057 scale points higher on items capturing *Ambitious Instruction* than 1999 lessons, a difference of approximately 0.3 standard deviations. The samples were indistinguishable in terms of the number of teachers' mathematical errors and imprecisions.

Table 5 shows in more detail the patterns that led to these findings. Lessons in our 2016 sample featured significantly more linking between representations, more mathematical sense-making, and more dense use of mathematical terminology, but no detectable difference in three other richness codes, including mathematical explanations, the use of multiple solution methods, and noticing patterns/building generalizations. Notably, these practices were the least frequent in both samples. Similarly, 2016 lessons featured more of some *Common Core-Aligned Student Practices*, including more student explanations, more student communication about mathematics, and more demanding contextualized problems. However, the two samples saw similar levels of task cognitive demand and student mathematical reasoning, and the samples did not differ in the extent to which teachers took up their students' mathematical ideas or offered strong remediation of student errors.

Finally, Table 5 shows one statistically significant difference in favor of the TIMSS sample lessons, with teachers in that sample solving fewer problems incorrectly. However, perhaps because the two samples did not differ in the level of precision in teachers' mathematical terms or the clarity of their mathematical talk, we observed no differences in averaged mathematical errors across samples (Table 4). The samples did not differ on *Overall MQI*, either; however, at the lesson level, our study was only powered to detect fairly large differences of at least 0.42 standard deviations.

Table 4
Average differences in MQI scores between 2016 and 1999, MQI factor composites

	Δ between 1999 and 2016	
Ambitious instruction	0.057	**
	(0.027)	
Errors	-0.024	
	(0.023)	
Number of segments	2,393	

Note. Standard errors provided in partentheses.
**p<.05

Table 5
Average differences in MQI item scores between 2016 and 1999

	Δ between 1999 and 2016	
<u>Richness</u>		
Linking Between Representations	0.097	***
	(0.037)	

Explanations	-0.016	
	(0.026)	
Math Sense-Making	0.143	***
	(0.066)	
Multiple Solution Methods	0.013	
	(0.039)	
Patterns and Generalizations	0.008	
	(0.020)	
Math Language	0.175	**
	(0.072)	
Working with students		
Remediation of Student Errors	-0.044	
	(0.055)	
Teachers Uses Student Contributions	0.039	
	(0.059)	
Errors		
Math Content Errors	-0.069	***
	(0.023)	
Imprecision in Language	0.038	
	(0.039)	
Lack of Clarity	-0.040	
	(0.029)	
Common-core aligned student practices		
Student Explanations	0.054	**
	(0.026)	
Student Reasoning	-0.023	
	(0.027)	
Student Communication	0.149	**
	(0.067)	
Task Cognitive Demand	0.010	
	(0.063)	
Contextualized Problems	0.150	**
	(0.070)	
Number of segments	2393	

Note. Standard errors provided in parentheses.
**p<.05 ***p<.01

RQ4: What instructional resources characterize 2016 lessons with higher instructional quality?

Curriculum is one key instructional resource. Table 6 shows where 2016 teachers reported finding or creating the lessons we observed on video. Notably, just over half of lessons (51%) did not rely upon a published textbook. Rather, in line with evidence from our nationally representative companion survey (Hill, Lovison & Kelly-Kemple, 2019), we see that teachers relied upon a range of resources, including internet repositories (24%) and supplemental materials (18%). Fifty-four percent of teachers said they created the observed lesson themselves and 29% said they created the lesson with colleagues. These numbers sum to more than 100% because teachers could report that they used several sources in developing their lesson, consistent with expectations that teachers might use different resources for different portions of their lesson.

Table 6
Teacher-reported origin of observed lesson

	Mean
Published textbook	0.49
State, district, or charter produced	0.06
Lesson created with colleagues	0.29
Lesson created by myself	0.54
Lesson obtained from internet	0.24
Released test item	0.05
Online content video	0.11
Published supplemental materials	0.18
Lessons	93

The top panel in Table 7 presents suggestive evidence that classroom instruction is stronger among teachers who used a textbook to plan their lesson. On average, *Overall MQI* scores were 0.24 points higher for teachers who used a textbook relative to those who did not (p = .09). Given our small sample size, and the substantive size of this difference – about one-third of a standard deviation – we consider this worthy of mention.The two right-hand columns in the top of Table 7 show that this relationship appears driven by a relationship between textbook use and *Richness* rather than *Errors*. Aside from textbooks, use of the other types of curriculum materials in Table 6 did not predict instructional quality.

Another instructional resource is teachers' mathematical knowledge. We could not measure teacher MKT in the context of this study, but our survey did ask teachers about their mathematical background. The bottom panel in Table 7 shows that teachers who reported a mathematics major, minor or graduate degree did not have substantially different *Overall MQI* scores than those who entered the profession with a more general background. However, teachers in our sample with math degrees did produce stronger lessons on the *Richness* dimension.

Table 7

	Number of Lessons	Overall MQI Score	Average Richness Score	Average Error Score
Lesson did not draw on textbook	47	2.77	1.51	1.12
		(0.76)	(0.30)	(0.28)
Lesson did draw on textbook	46	3.01	1.56	1.13
		(0.62)	(0.28)	0.27
Difference	93	-0.24*	-0.05**	-0.00
		(0.14)	(0.02)	(0.02)
Teacher did not have math degree	59	2.86	1.51	1.14
		(0.62)	(0.28)	(0.29)
Teacher did have math degree	34	2.94	1.56	1.11
		(0.82)	(0.32)	(0.27)
Difference	93	-0.09	-0.05**	0.03*
		(0.15)	(0.02)	(0.02)

Notes: In each panel, rows one and two of columns 2-4 report means with standard deviations in parentheses. Row 3 of each panel reports the results of a two-sample t-test. Standard errors reported in parentheses.*p<.10, **p<.05.

Discussion and conclusion

This study asked about the prevalence of ambitious instruction in U.S. middle school mathematics classrooms using the MQI instrument to gauge this construct. Doing so allowed us to take stock of mathematics instruction several decades into a concerted effort to transform that instruction, and six years after the Common Core State Standards. Our sample also allowed us to observe correlates of stronger instruction for a limited set of variables.

Our primary results suggest strong variability in the quality of U.S. middle school mathematics instruction. Using scores from *Overall MQI*, we estimate that roughly three in ten lessons exhibited a fair amount of ambitious instruction, another five in ten lessons exhibited either more traditional instruction or a mix of some stronger and weaker elements, and two in ten lessons contained instruction our raters marked as concerning. This distribution differs from accounts of mathematics instruction pre-reform, which suggested little attention to sense-making and mathematical reasoning, and that students' roles were limited to learning and practicing procedures.

Qualitative analyses of a subsample of lessons also illustrate the variability in students' experiences in these classrooms. Among higher-scored lessons, teachers invited mathematical sense-making and student participation in the mathematics. They did so, however, within a traditional, teacher-centered instructional format. Our lowest-scored lessons featured either minor to moderate mathematical errors and/or confusing instruction, enough that students may have had difficulty learning. Lessons scored closer to the center of the distribution featured brief and infrequent standards-aligned practices, and were again enacted within the confines of teacher-centered instruction. These average-quality lessons featured many activities likely not found in classrooms decades ago (e.g., Barbie Bungee Jumping), but their enactment did not reach the level of mathematical depth and rigor lesson authors intended.

Our analyses of these lessons were also notable for what we did not see. We saw no discussion-based mathematics lessons. We saw little student work on mathematically novel or challenging problems. We saw little in the way of argumentation, proof, justification, or other mathematical practices. Though our sample was small, we saw no classrooms of the kind featured in case studies of successful standards-aligned classrooms. This suggests that at best, changes in U.S. instruction have been incremental rather than transformational.

All our results must be interpreted cautiously in light of our response rate, which is low by industry standards and in comparison to recent district-specific studies (e.g., Hill, Litke & Lynch, 2019; Kane & Staiger, 2012). However, there is some indication that our teacher sample did not differ significantly from a higher-response nationally representative sample in terms of mathematics and educational coursework; if anything, the video sample teachers were slightly less prepared, in terms of undergraduate mathematics degrees, which runs counter to the expectation that more expert teachers would respond to the video study.

Comparing results from our 2016 sample to other studies that have used the MQI to score classroom observation data can further help illuminate the degree of self-selection in our sample. Hill, Litke & Lynch (2019) analyzed instruction in five districts and found modestly more disciplinary richness and student thinking and reasoning than in our sample, but their sample included two districts using *Investigations*, a set of curriculum materials that supported reform-aligned instruction. By contrast, the Measures of Effective Teaching study (Kane & Staiger, 2012), which also included five districts, found very little richness or student participation in mathematical meaning-making and reasoning. Thus our results seem to fall in between these two studies that featured only a small number of districts.

Differences in sample response rates also complicate comparisons between the 1999 (TIMSS) and 2016 samples, and our results must be considered in light of these challenges. Most generally, we found modest differences in the mathematical quality of instruction across these two time points. Lessons captured in 2016 contained more *Richness* elements, and more student explanation and communication. However, the cognitive demand of the average task did not differ, nor did teachers' propensity to take up and use students' contributions, or to remediate student misunderstandings at a conceptual level. Assuming no positive selection (i.e., stronger teachers more likely to participate) in the 2016 sample, this suggests incremental change towards ambitious instruction. Assuming modest positive selection into the 2016 sample, this suggests that instruction has remained largely unchanged since the late 1990s.

We next examined whether we could identify correlates of lesson quality. Because of questions about the quality of materials used to enact instruction, we asked teachers to describe the origin of the lesson we captured. Similar to a sister project (Hill, Kelly-Kemple & Lovison, 2019), we found that a slight majority of lessons in our sample did not use a published textbook. We also found that lessons based on textbooks outperformed lessons based on other resources, as measured by *Overall MQI*. This pattern was particularly true for *Richness* codes, suggesting that lessons gleaned from other sources and/or created by teachers themselves may lack support for high-quality teaching. We also found a relationship between *Richness* scores and teachers' possession of a mathematics degree. This suggests that policies such as No Child Left Behind, which attempted to require more mathematical coursework for middle school teachers, could be beneficial to classrooms.

Our small study made several other observations of interest. First, although we used a different instrument than TIMSS to score the 1999 video data, our results appear quite consonant, substantively, with those reported in Hiebert et al. (2005). Notably, even in cases where teachers posed students a mathematically rich and challenging mathematical task, they ultimately devolved that task to lower levels of cognitive demand by the end of the lesson. And our results are consonant with another study (Hill, Litke & Lynch, 2019) that showed even in classrooms that came closer to reformers' vision, teachers generally accomplished this goal by interspersing moments of cognitive demand and mathematical meaning into more traditional, teacher-centered instruction.

In sum, results from our 2016 sample and similar evidence suggests that mathematics instruction in the U.S. does feature some of the characteristics reformers desire, but is far from being characterized by those features. One reason for this modest progress may be the uneven national approach to standards-aligned instruction. Although most official state standards either currently or in the recent past resemble the Common Core and its predecessors, standards exist amidst an array of other policy initaitives and guidance. The same decades that saw a push for standards-aligned mathematics instruction also saw strong efforts to implement test-based accountability at both the school and teacher level. Schools and teachers responding to the latter effort would prioritize improvement efforts that led to enhanced student performance on state standardized tests. Scholars have argued that such assessments may not be sensitive to standards-aligned instructional practice (e.g., Sussman & Wilson, 2018), meaning that test-based accountability policies might dilute attention to achieving ambitious instruction. Further, schools and teachers who intended to implement standards may have found uneven resources for doing so (Kennedy, 2005) – curriculum materials that claim to be standards-aligned but were only partially so, professional development of low quality, and so forth.

A second reason for the modest presence of ambitious instruction 2016 may be the enduring appeal of direct instruction in U.S. classrooms. As David Cohen (2012) notes, teacher-centered instruction eases the cognitive challenge of providing instruction by allowing only limited student input, thus reducing the scope of ideas in play and the number of decisions that teachers must make (see also Kennedy, 2005). What we observed in classrooms – teachers mixing some ambitious practices into otherwise conventional instruction – would fit this storyline, as teachers retain control of classroom discourse while adding moments of heightened student thinking and mathematical meaning.

Whatever the case, progress toward ambitious instruction has been slow. If the U.S. is to commit to more intellectually rigorous instruction, it needs to reconsider its strategy, possibly by supporting teachers to use high-quality, standards-aligned materials, and probably by bringing policy instruments into alignment with those standards and materialss. Reformers may also want to re-engineer their change efforts to take into account the fact that even the strongest lessons in our sample used traditional structures for instruction. Rather than subverting these structures, leveraging change within these structures may pose another avenue forward.

References

Ball, D.L., (1990). Reflections and deflections of policy: The case of Carol Turner. *Educational Evaluation and Policy Analysis, 12*(3), 247-259.

Ball, D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of teacher education*, *59*(5), 389-407.

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M. L. (2018). Report of the 2018 NSSME+. *Horizon Research, Inc.*

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn* (Vol. 11). Washington, DC: National Academy Press.

California State Department of Education (1985). Mathematics curriculum framework for California public schools. Sacramento, CA. Retrieved from http://www.cde.ca.gov/ci/cr/cf/documents/mathfrwk.pdf#search=mathematics%20framework%201985&view=FitH&pagemode=none.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*, (3), 311-330.

Cohen, D. K. (2011). *Teaching and its predicaments*. Harvard University Press.

Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis*, *12*(3), 331-338.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., ... & Doolittle, F. (2011). Middle School Mathematics Professional Development Impact Study: Findings after the Second Year of Implementation. NCEE 2011-4024. *National Center for Education Evaluation and Regional Assistance*.

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., ... & Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. *Educational Evaluation and Policy Analysis*, *27*(2), 111-132.

Hill, H.C., Kapitula, L.R. & Umland, K. L (2011). A validity argument approach to evaluating value-added scores. *American Educational Research Journal 48(3),* 794-831.

Hill, H.C., Lovison, V.S. & Kelley-Kemple, T. (in press). Mathematics Teacher and Curriculum Quality, 2005 and 2016. *AERAOpen.*

Hill, H.C., Umland, K. L., Litke, E. & Kapitula, L. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education, 118, 489-519.*

Jacobs, J., Garnier, H., Gallimore, R., Hollingsworth, H., Givvin, K. B., Rust, K., ... & Etterbeek, W. (2003). Third International Mathematics and Science Study 1999 Video Study Technical Report: Volume 1--Mathematics. Technical Report. NCES 2003-012. *National Center for Education Statistics*.

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for research in Mathematics Education*, *38*(3), 258-288.

Johnson, T. P., & Wislar, J. S. (2012). Response rates and nonresponse errors in surveys. *JAMA*, *307*(17), 1805-1806.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kelcey, B., Hill, H. C., & Chin, M. J. (2019). Teacher mathematical knowledge, instructional quality, and student outcomes: a multilevel quantile mediation analysis. *School Effectiveness and School Improvement*. DOI: 10.1080/09243453.2019.1570944

Kennedy, M. M., (2005). *Inside teaching: How classroom life undermines reform*. Cambridge, MA: Harvard University Press.

Lloyd, G. M., & Wilson, M. (1998). Supporting innovation: The impact of a teacher's conceptions of functions on his implementation of a reform curriculum. *Journal for research in mathematics education*, *29*(3), 248-274.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199-226.

National Council of Teachers of Mathematics. Commission on Standards for School Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics.*

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.*

National Research Council. (2013). *Monitoring progress toward successful K-12 STEM education: A nation advancing?* Washington, DC: The National Academies Press.

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). Common Core State Standards (Mathematics). *National Governors Association Center for Best Practices, Council of Chief State School Officers*.

O'Connor, C., Michaels, S., & Chapin, S. (2015). Scaling down" to explore the role of talk in learning: From district intervention to controlled classroom study. *Socializing intelligence through academic talk and dialogue*, 111-126.

Porter, A. (1989). A curriculum out of balance the case of elementary school mathematics. *Educational Researcher*, *18*(5), 9-15.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., ... & Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, *47*(4), 833-878.

Sussman, J., & Wilson, M. R. (2018). The Use and Validity of Standardized Achievement Tests for Evaluating New Curricular Interventions in Mathematics and Science. American Journal of Evaluation, 1098214018767313.

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). A study of K-12 mathematics and science education in the United States. *Chapel Hill, NC: Horizon Research*.

Wilson, S.M., (1990) A Conflict of Interests: The Case of Mark Black. *Educational Evaluation and Policy Analysis*, *12*(3) (Autumn, 1990), pp. 293-310.

Zahs, D. Gilbert, B. & Herlihy, C. (2018) MTTS data book. Cambridge, MA: Authors.