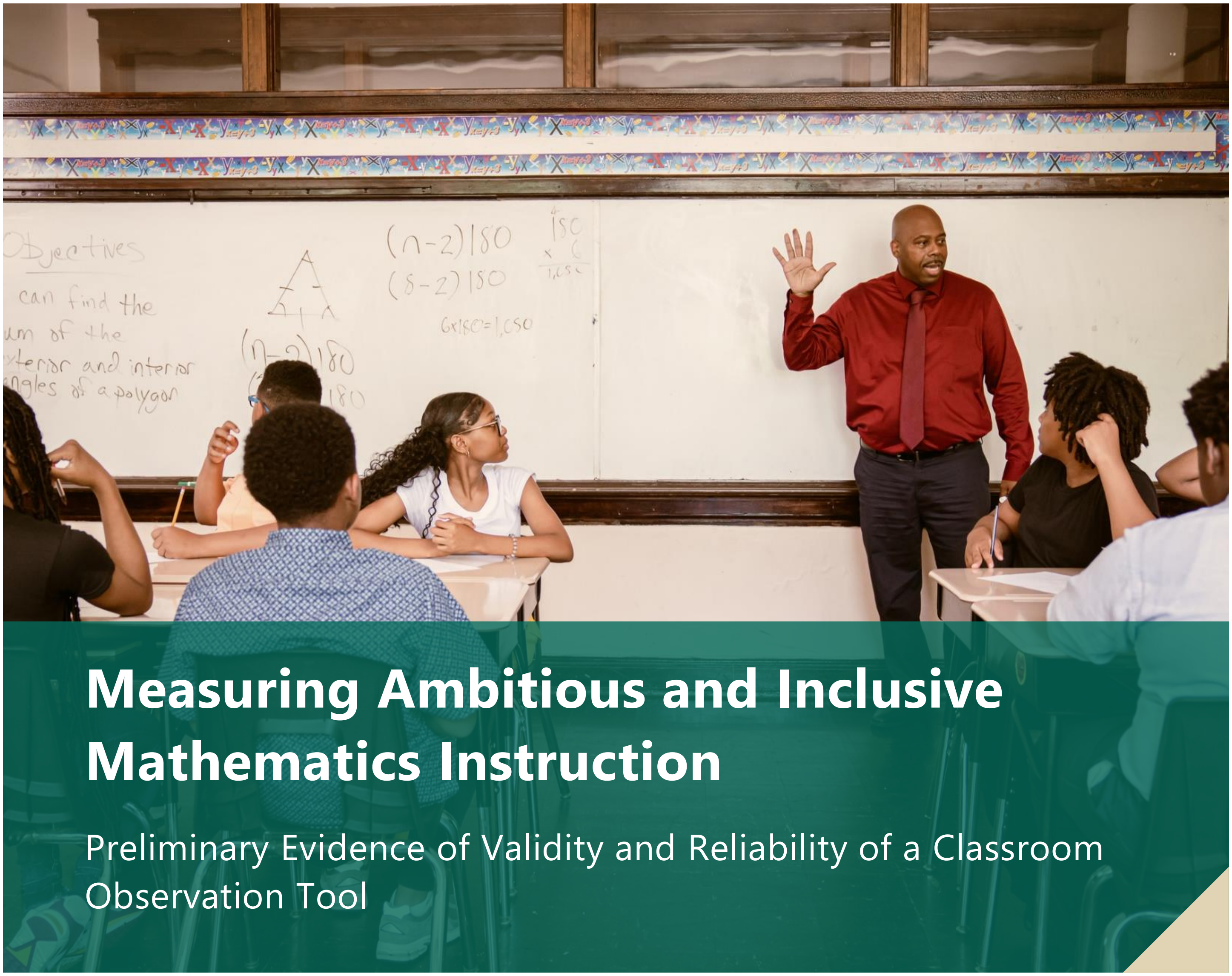


Measuring Ambitious and Inclusive Mathematics Instruction

Preliminary Evidence of Validity and Reliability of a Classroom Observation Tool

December 31, 2024

Lauren Amos and Micah Wood

Submitted to:
Bill & Melinda Gates Foundation
Attention: Andy Sokatch

Submitted by:
Mathematica
P.O. Box 2393
Princeton, NJ 08543-2393
Phone: (609) 799-3535
Fax: (609) 799-0005

ACKNOWLEDGEMENTS

We greatly appreciate the contributions of the many people who have helped make the AIM classroom observation tool and this methodological report possible. First and foremost, we express thanks for our study co-principal investigators:

Laura Desimone, Ph.D.
Professor and Director of Research in the College of Education and Human Development
University of Delaware

Maria Zavala, Ph.D.
Associate Professor of Elementary Education
San Francisco State University, Graduate College of Education

Robert Sheffield, Ed.D.
Former Director of Curriculum, Assessment, and Instruction Services
WestEd

We are deeply grateful for the extensive support of Drs. Temple Walkowiak and Holly Pinter in assessing the convergent and divergent validity of the AIM observation tool with the Mathematics Scan measure.

Temple A. Walkowiak, Ph.D.
Associate Professor, Mathematics Education
University Faculty Scholar
Department of Teacher Education & Learning Sciences
North Carolina State University

Holly Henderson Pinter, Ph.D.
Associate Professor, Western Carolina University
Program Coordinator, BSED: Elementary & Middle Grades
Catamount School Instructional Support and University Liaison
School of Teaching and Learning

We also thank the Analysis of Middle School Math System's Math Advisory Council for its guidance and recommendations as we developed early iterations of the tool. Members of the council include:

Benjamin Aguilar
Professional Learning Manager
Open Up Resources

Nathan Alexander, Ph.D.
James King Jr. Professor of Mathematics Teaching and Associate Director, Communicating by Thinking Effectively in and About Mathematics
Morehouse College

Robert Q. Berry III Ph.D.
Samuel Braley Gray Professor of Mathematics Education and Associate Dean of Diversity, Equity, and Inclusion
University of Virginia

Doug Jaffe, J.D.
Former Senior Director, Sustainability, Office of Portfolio Development
New York City Department of Education

Sarah Johnson, Ed.D.
Chief Executive Officer
Teaching Lab

Nicole Joseph, Ph.D.
Assistant Professor of Mathematics Education, Department of Teaching and Learning
Vanderbilt University

David Kirkland, Ph.D.
Associate Professor of English and Urban Education and Executive Director
New York University Metropolitan Center for Research on Equity and the Transformation of Schools

Jeff Livingston
Chief Executive Officer
EdSolutions

Aly Martinez
Former Instructional Coordinator TK-12 Mathematics
San Diego Unified School District

Lisette Nieves, Ed.D.
Director of Educational Leadership and full Clinical Professor of Educational Leadership and Policy Studies
NYU Steinhardt

John Staley, Ph.D.
Past Chair
U.S. National Commission on Mathematics Instruction
Past President
NCSM
Former Director Mathematics
Baltimore County Public Schools

Julie Washington, Ph.D.
Professor and Chair of the Department of Communication Sciences and Disorders
Georgia State University

Jason Zimba, Ph.D.
Founding Partner
Student Achievement Partners

Abstract

This paper introduces and explores the validity and reliability of a classroom observation tool that we developed to measure **ambitious** (cognitively demanding and standards-based) and **inclusive** (culturally responsive, linguistically responsive, and equitable) mathematics teaching. The tool was developed to inform a multiyear study of the enactment of middle school math curricula in four urban school districts. We begin by defining ambitious and inclusive instruction and present an overview of existing observational measures designed to measure ambitious or inclusive practice. Second, we discuss the iterative design and pilot test of our classroom observation tool, including how we assessed *content validity* by expert review, evaluated the tool's *internal consistency* using Cronbach's alpha, and assessed *interrater reliability* using Cohen's kappa. Third, we share the results of two tests of validity: (1) *convergent and discriminant validity* using an existing observational measure of ambitious practice and (2) *construct validity* using student survey data to assess the extent to which our tool can measure inclusive practice. Results indicate that the Ambitious and Inclusive Mathematics (AIM) classroom observation tool is a promising measure of cognitively demanding, standards-based mathematics instruction that is culturally responsive, linguistically responsive, and equitable. We found that employing ambitious and inclusive practices is positively associated with non-academic student outcomes such as math enjoyment, engagement, math achievement identity, self-efficacy, and growth mindset. In addition, we affirmed our assertion that inclusive practices are inherently ambitious. The use of inclusive practices should not be regarded as academic enrichment or supplemental and should not supplant a focus on rigor or learning standards. They should be employed in an integrated manner to improve student outcomes. Inclusive practice serves all students regardless of their race, ethnicity, linguistic traditions, or cultural heritage.

Table of contents

Exhibits

I. Introduction

The COVID-19 pandemic disrupted learning for millions of students. Historically marginalized communities and high-poverty schools were particularly hard hit (Nowicki, 2022). As districts and schools weighed strategies to mitigate adverse impacts of the pandemic on students' social-emotional well-being and academic experiences, addressing the needs of students who are Black, Latino, multilingual learners, or experiencing poverty was a high priority for many US public school systems. To better meet the needs of these students, building teacher capacity to engage in *inclusive* pedagogical practice—culturally and linguistically responsive and equitable—while holding students to high expectations with *ambitious* instruction—standards-based and cognitively demanding—was often the focus of district and school level investments in professional learning, remote and hybrid learning, and instructional materials.

Box A.1. Key definitions and concepts

The mathematics education practitioner and research communities define ambitious and inclusive instruction in various ways. This project and report use the following definitions.

- **Ambitious instruction:** We refer to *ambitious* practice as cognitively demanding, standards-based instruction (Jackson & Cobb, 2010; Kazemi et al., 2009; Stroupe, 2016).
- **Inclusive instruction:** We refer to *inclusive* practice as instruction that is culturally and linguistically responsive as well as equitable.
- *Culturally responsive instruction:* We draw on research conducted by Aguirre and del Rosario Zavala (2013), who define culturally responsive mathematics teaching as pedagogical knowledge, beliefs, dispositions, student expectations, and practices that collectively promote mathematical thinking, use cultural and linguistic funds of knowledge as an instructional asset, and employ mathematics as a tool for social justice (Aguirre & del Rosario Zavala, 2013; Jones, 2015; Moll et al., 2006; Turner et al., 2012).
  - *Linguistically responsive instruction:* We refer to the use of English-language scaffolding strategies or providing translation support to make a math-related conversation or task more accessible to multilingual learners (Aguirre & del Rosario Zavala, 2013; Hanzlian, 2013; Jones, 2015; Turner et al., 2012; Civil, 2016; National Academies of Sciences, Engineering, and Medicine, 2018; Erath et al., 2021; Moschkovich, 2013; Moshckovich, 2015; de Araujo et al., 2018).
  - *Equitable instruction:* We define equitable teaching as instructional protocols, tasks, or content that personalizes or differentiates the learning experience for specific subgroups of students, such as multilingual learners and students with disabilities, to ensure that all students have equal access and opportunity to engage in the learning process. Equitable teaching practices such as wait time are not necessarily culturally or linguistically responsive.

Although we define ambitious and inclusive instruction separately, our work rests on the belief that **inclusive practices are inherently ambitious**. Contrary to dynamics in the education community that have marginalized culturally responsive education (Aronson & Laughter, 2016), **the use of inclusive practices should not be regarded as academic enrichment or supplemental and should not supplant a focus on rigor or learning standards.** They should be employed in an integrated manner to improve student outcomes. Inclusive practice serves all students regardless of their race, ethnicity, linguistic traditions, or cultural heritage (Sleeter, 2012).

With this context in mind, the Bill & Melinda Gates Foundation's Coherent Instructional Systems investment portfolio is grounded in the belief that students who are Black, Latino, multilingual learners, or experiencing poverty will succeed when they are served within a **coherent instructional system**. A coherent system is one in which district and school visions for high-quality instruction are aligned with the provision of high-quality standards-based curricula, effective professional learning, and instructional practice that is ambitious and inclusive.

Under this portfolio, Gates partnered with Mathematica to conduct the **Analysis of Middle School Math Systems (AMS) study** to investigate the enabling and disabling conditions (such as access to a math coach, curriculum-specific professional learning, collaborative planning time, and supportive leadership) under which teachers adopted and adapted six different middle school mathematics curricula in four urban school districts: Illustrative Math, Into Math, Eureka Math, California Math, Big Ideas, or Key Elements of Mathematics Success (KEMS). In addition to conducting extensive interviews with district and school staff, administering teacher and student surveys, conducting student focus groups, and observing professional learning and coaching sessions, we conducted nearly 90 classroom observations over the course of two consecutive school years to better understand how the instructional resources and supports teachers receive at the district and school levels influence the extent to which teachers create more ambitious and inclusive learning environments for their students.

Middle school is a critical time during which students begin to make decisions about whether to pursue college preparatory coursework in mathematics in high school. Race, ethnicity, and poverty are among the most significant predictors of rigorous mathematics course taking (Sciarra, 2010). Black and Latino students, particularly those experiencing poverty, are less likely to enroll in cognitively demanding mathematics courses in secondary school (Riegle-Crumb & Grodsky, 2010). Even when controlling for prior achievement, mathematics course-taking patterns play a critical role in explaining variations in academic performance outcomes (Wang & Goldschmidt, 2003), and failing a mathematics course in middle school is a stronger predictor of not graduating from secondary school than are low test scores (Balfanz et al., 2007). Students who do not believe they can perform well in mathematics tend to perform at lower levels than students who believe they can excel (see, for example, Chen, 2003; Cleary & Chen, 2009; Goetz et al., 2008; Lopez, 2017; Mason & Scrivani, 2004; Pinxten et al., 2014; Riegle-Crumb et al., 2011; Schommer-Aikins et al., 2005). Consequently, our study also explored the extent to which ambitious and inclusive practices can positively influence students' experiences in math classrooms, such as their growth mindset, math identity, persistence, enjoyment, self-efficacy, and engagement.

Research questions

To explore the nature and role of ambitious and inclusive instruction in fostering positive student experiences, we conducted observations of middle school mathematics classrooms—predominantly serving Black students, Latino students, or students experiencing poverty—in four urban school districts over the course of two consecutive school years following COVID-related school closures. We investigated the extent to which and how ambitious and inclusive practices foster positive student experiences in middle school math classrooms.

This paper introduces and explores the validity and reliability of the Ambitious and Inclusive Mathematics (AIM) classroom observation tool that Mathematica developed to help identify pedagogical approaches that foster positive student experiences in math. Specifically, we ask the following questions about the AIM tool:

- Is it **reliable and valid for use in a large-scale study across multiple contexts** (curricula, districts, schools, classrooms, and instructional units)?
- Is it **empirically supported**? Do data collected by the AIM tool:
  – Support the assertion that inclusive practices are positively associated with ambitious practices?
  – Demonstrate that inclusive practice positively influences student belief (math enjoyment, math achievement identity, math self-efficacy, and growth mindset) and engagement in math?
  – Affirm our hypothesis that procedural learning environments are less ambitious and can negatively influence student belief and engagement in math?

We begin by detailing why we chose to develop a new tool rather than use an existing instrument. In the section that follows, we detail the design and development of the AIM classroom observation tool. We then discuss our approach to testing the psychometric properties of the tool and present the results of these analyses. We close with a discussion of our study limitations and next steps.

Existing measures of ambitious and inclusive instruction

The research record on culturally and linguistically responsive and equitable instructional practice is inspirational but largely qualitative, theoretical, anecdotal, and aspirational. There is a dearth of actionable, scalable, and causal research illustrating effective implementation of these strategies and evidencing that the use of these practices contributes to improved student outcomes. A review of culturally responsive measures found that the majority were teacher self-report surveys, with few drawing on student reports or assessments by external observers (Franco et al., 2024). At the outset of the study, we conducted a literature review and landscape analysis to identify classroom observation instruments that meet the following requirements:

- They are **reliable and valid** for use in a large-scale study:
  – *Across multiple contexts* (curricula, districts, schools, classrooms, and instructional units)
  – For both video-recorded and in-person lessons
- They are appropriate or adaptable for assessing instruction in **middle school math learning environments.**
- They can be used to document both **culturally responsive** teaching AND **equitable** teaching (inclusive instructional strategies intended to differentiate or personalize instructional content and tasks to ensure all students have equal access to the learning experience, such as heterogenous and cooperative groupings). Our landscape analysis indicated that many instructional resources promote teaching practices that are equitable (such as "wait time," where teachers pause conversation long enough for students to collect their thoughts and respond to a question or prompt) that are not necessarily culturally or linguistically responsive.
- They score the occurrence or non-occurrence of **observed behavior, activity, or speech** rather than require a coder to inferentially *evaluate or rate the quality of observed behavior* so that the tool can be used reliably by researchers who do not have math education expertise or substantial teaching experience in math.

- They are **empirically supported** by culturally responsive practices that have been documented in research on effective or promising practice, rather than *aspirational or theoretical* approaches presenting an ideal or vision for culturally responsive practice.

We identified and reviewed nine existing tools: (1) Reform-Oriented Teaching Observation Protocol (RTOP; Sawada et al., 2002; Boston et al., 2015); (2) Instructional Quality Assessment in Mathematics (IQA; Boston & Candela, 2018; Boston et al., 2015); (3) Mathematical Quality of Instruction (MQI; Learning Mathematics for Teaching Project, 2011; Boston et al., 2015); (4) Comprehensive Mathematics Instruction Observation Protocol (CMI; Womack, 2011); (5) Electronic Quality of Inquiry Protocol (EQUIP; Marshall et al., 2010); (6) Mathematics Scan (M-Scan; Walkowiak et al., 2014); (7) Assessing Classroom Sociocultural Equity Scale (ACSES; Curenton et al., 2019); (8) systematic approach to culturally responsive practices across classrooms (CRP; Larios et al., 2022); and (9) Culturally Responsive Instruction Observation Protocol (CRIOP; Powell et al., 2016; Powell et al., 2013). Ultimately, we were unable to identify a classroom observation instrument that met all of our criteria. Across these nine tools, five (IQA, MQI, CMI, EQUIP, and M-Scan) were math-specific but did not include measures or components to observe culturally or linguistically responsive practice. The three tools that did measure both culturally responsive and equitable practice (ACSES, CRP, and CRIOP) were not explicitly designed for math environments. The ninth tool (RTOP) satisfied neither of these criteria. In additions, only four of the nine tools (RTOP, MQI, ACES, and CRIOP) were scored deductively; the majority relied on a determination of evaluated behavior based on a rubric or set of standards. The characteristics of the nine existing tools we reviewed are summarized in Exhibit I.1. Descriptions of each, including how each did not meet our selection criteria, are available in Appendix A.

Exhibit I.1. Comparison of existing observation tools

Tools	Reliable and valid for use in a large-scale study	Can be used across multiple contexts and curriculums	Can be used with recorded lessons	Math-specific	Middle school-specific	Measures culturally responsive	Measures equitable practice	Scored deductively	Empirically supported
RTOP	✓	✓	✓	X	X	X	X	✓	?
IQA	✓	✓	✓	✓	X	X	X	X	?
MQI	✓	✓	✓	✓	X	X	X	✓	✓
CMI	?	X	?	✓	X	X	✓	X	X
EQUIP	✓	✓	?	✓	X	X	X	X	✓
M-Scan	✓	✓	✓	✓	✓	X	✓	X	✓
ACSES	✓	✓	✓	X	X	✓	✓	✓	✓
CRP	✓	✓	✓	X	✓	✓	✓	X	✓
CRIOP	✓	✓	?	X	X	✓	✓	✓	✓
AIM	→	✓	✓	✓	✓	✓	✓	✓	✓

Note: A check mark indicates that the measure contains or demonstrates this characteristic; an X indicates it does not; a question mark indicates we could not determine this based on the information provided in the manuscript. An arrow indicates there is preliminary evidence of reliability and validity.

II. The Ambitious and Inclusive Mathematics (AIM) Classroom Observation Tool

Based on our review of existing observational tools, we elected to develop our own tool. We iteratively developed the Ambitious and Inclusive Mathematics (AIM) classroom observation tool over a four-year period. In the fourth year, we tested the tool's psychometric properties and refined its final design based on the tool's validity results. In this section, we discuss how we designed the tool; how we used the tool to observe instructional practice; how we summarize, interpret, and report data generated by the tool; and how we trained and certified coders to use the tool.

AIM tool design and development

To develop the initial version of the AIM tool, we adapted Aguirre and del Rosario Zavala's (2013) culturally responsive mathematics teaching (CRMT) lesson analysis tool. The CRMT lesson analysis tool is intended for teachers to use in a professional learning setting to self-reflect on their practice in nine areas prompted by a set of guiding questions:

1. Intellectual support
2. Depth of student knowledge and understanding
3. Mathematical analysis
4. Mathematical discourse and communication
5. Student engagement
6. Academic language support for multilingual learners
7. Use of English as a second language (ESL) scaffolding strategies
8. Funds of knowledge/culture/community support
9. Use of critical knowledge/social justice

We built on these dimensions to design a tool that would be appropriate for researchers to use when observing a teacher, as well as to collect detailed data about the learning environment that the CMRT was not designed to systematically document (such as the range of student grouping strategies used and cognitive demand of student performance tasks assigned during a lesson). In addition, we drew on (1) research on culturally responsive teaching, multilingual learning, and equitable practice in mathematics; (2) the National Council of Teachers of Mathematics Principles to Action; and (3) recommendations and feedback from the study's Math Advisory Council, comprising experts in mathematics education, professional learning, middle grades teaching and learning, and culturally responsive pedagogy.

The AIM observation tool documents the extent to which teachers employ the instructional practices outlined in Exhibit II.1.

Exhibit II.1. Core AIM instructional practice domains

Domain	Description	Example instructional practice
Real-world mathematical inquiry and problem solving	Mathematics instruction that explicitly requires students to pose questions and investigate authentic problems (Aguirre & del Rosario Zavala, 2013; Jones, 2015; Turner et al., 2012)	Teacher poses a mathematical question, problem, or task with explicit real-world implications or poses a mathematical question, problem, or task that requires applying real-world data or information to solve
Multiple representations of mathematics	Mathematics instruction that encourages multiple ways of knowing and expressing mathematical ideas and embraces multiple solution paths (Ainsworth, 2006; Edmonds-Wathen, 2019; Jitendra et al., 2007; Pape & Tchoshanov, 2001; Selling, 2016)	Teacher encourages students to share, discuss, or demonstrate: • Their reasoning and sense making about different symbolic, textual, or graphical representations of mathematical concepts or relationships • Connections or relationships of the mathematical concepts, procedures, or tasks at hand with other mathematical ideas (e.g., presented in a different lesson) • Alternative solution paths
Mathematical discourse	Mathematics instruction that creates opportunities for students to discuss mathematics in meaningful and rigorous ways (e.g., debate mathematics ideas/solution strategies, use mathematics terminology, develop explanations, communicate reasoning, or make generalizations) (Aguirre & del Rosario Zavala, 2013; Jones, 2015; Turner et al., 2012)	Teacher probes or asks purposeful questions, or provides instructions to engage more than one student to: • Evaluate or compare each other's representations, solutions, approaches, or arguments • Debate math ideas and strategies • Co-construct strategies or explanations in response to a mathematical task
Multilingual learner language support and scaffolding	Mathematics instruction that draws on multiple modes of communication and regards students' home languages as instructional resources, rather than a deficit, to support the academic language development of multilingual students regardless of their English proficiency (Aguirre & del Rosario Zavala, 2013; Hanzlian, 2013; Jones, 2015; Turner et al., 2019; Civil, 2016; National Academies of Sciences, Engineering, and Medicine, 2018; Erath et al., 2021; Moschkovich, 2013; Moshckovich, 2015; de Araujo et al., 2018)	Teacher uses an English-language scaffolding strategy or provides translation support to make a math-related conversation or task more accessible

Domain	Description	Example instructional practice
Engaged student and community funds of knowledge	Mathematics instruction that helps students bring their lived experience and intuitive knowledge to the instructional setting as an asset for individual and collective learning (Aguirre & del Rosario Zavala, 2013; Jones, 2015; Moll et al., 2006; Turner et al., 2012)	Teacher connects or employs students' community, cultural, or linguistic knowledge that is specific to their individual lived experience or local context with a math-related discussion or task
Interdisciplinary connections	Mathematics instruction that draws on connections from other content areas and domains of study (Aguirre & del Rosario Zavala, 2013; Jones, 2015; Turner et al., 2012)	Teacher explicitly connects a math-related discussion or task to another academic discipline or content area (e.g., science, social studies, or art) as a tool to broaden students' understanding and application of a mathematical fact, concept, or procedure beyond the lesson
Empowered mathematical inquiry and decision making	Mathematics instruction that engages students in posing questions about societal challenges of relevance to them and tasks that explore, critique, and posit or test solutions to those issues (Aguirre & del Rosario Zavala, 2013; Jones, 2015; Turner et al., 2012)	Teacher poses a question, initiates a discussion, or assigns an instructional task that requires students to use math to investigate or critique a societal challenge or a social justice issue of direct relevance to them or of their own choosing

In addition, the tool documents characteristics of the learning environment in which these practices are (or are not) employed, including the following:

- Teachers' use of **student grouping strategies** (such as when students work in small groups or peer pairs)
- Student–teacher and student–student **relational interactions** (Battey et al., 2018) (such as instances when a teacher encourages a student to work through a difficult task or addresses off-task behavior)
- **Administrative procedures and classroom protocols** (routine- or protocol-driven tasks such as taking attendance and assigning homework)
- **Procedural instruction** (scripted or routine-driven instruction such as lecturing and administering exit tickets)
- The cognitive demand of the **performance tasks** teachers assign to students
- Teachers' use of **core and supplemental instructional materials** (such as educational technology, language aids for multilingual learners, and teacher-developed resources)

Coder training and certification

A team of Mathematica analysts with qualitative research or classroom teaching experience participated in a five-day training on the tool before conducting classroom observations with the tool each school year. Coders were education researchers, half of whom had math teaching experience. The training was co-designed and co-facilitated by the lead developer of the AIM classroom observation tool (the gold standard coder) and a senior member of the research team. The gold standard coder is a learning scientist with middle grades classroom teaching experience and expertise in culturally responsive teaching. The training involved the following:

- Group discussions of the research base on culturally responsive teaching and for each of the domains in the tool
- Group discussion of the codes in each domain, including their definition, as well as inclusion and exclusion criteria indicating when to apply a code
- Group coding practice using brief video examples of classroom practice
- Independent coding practice using longer video examples of classroom practice

Following training and before data collection began, coders received additional opportunities to independently practice using the tool with new video examples. To certify coders, the gold standard coder or the tool's co-developer (a senior member of the research team) tested trainees' independent coding practice for interrater reliability using Cohen's kappa (Cohen, 1960). Coders continued to practice using the tool until coder agreement with the gold standard exceeded 80 percent.

Throughout the training process, we refined the tool's codes, descriptions, inclusion and exclusion criteria, and examples based on coder feedback.

Coding procedure

The AIM tool was designed as a low-tech, Excel-based tool. This format enabled us to integrate the codebook into the tool so that coders could easily access code descriptions, coding inclusion and exclusion criteria, and examples. The digital format also made it possible to automate scoring observation data and auto-calculate interrater reliability (IRR). Using the AIM tool, coders observed and coded entire class periods (including those in which more than one lesson was delivered, such as during a 90-minute class period). In five-minute intervals, coders documented whether a specific practice or behavior occurred at least once during a five-minute interval. Observed class periods ranged in length from 35 to 90 minutes (7 to 18 intervals) with an average of 12 intervals (60 minutes). Some codes represent teacher behaviors and speech, whereas others represent student behaviors and speech, so that scores can distinguish between student and teacher participation patterns.

- **In-person observations** were coded by two certified coders. One was responsible for taking detailed notes on classroom activity in five-minute intervals, and the second coder was responsible for using the tool to code classroom activity in five-minute intervals. Coders were responsible for collecting or photographing instructional materials used during a lesson and content displayed or referenced by the teacher (such as on a whiteboard or transparency machine), to the extent feasible. Coders met within 24 hours after each observation to discuss and resolve coding questions based on the detailed notes, so that codes reflected a consensus between coders.
- **Video observations** were coded by one coder using a recording of an entire class period stored in IRIS Connect (https://www.irisconnect.com/us/), a secure, cloud-based, and customizable classroom video-recording and sharing platform used by educators and education researchers. IRIS Connect provides audiovisual recording kits to support data collection. Mathematica staff with significant experience conducting remote, virtual classroom observations recruited, trained, scheduled, and monitored field staff in three of the study districts to set up, record, and upload video recordings. Field staff were also responsible for collecting or photographing instructional materials used during a lesson and content displayed or referenced by the teacher (such as on a whiteboard or a chart on the classroom wall), to the extent feasible. Video coders met weekly as a team to discuss inclusion and exclusion criteria for specific codes, as questions arose. If coding criteria were clarified or refined based on these conversations, all coders revised previously coded observations to reflect changes to the codebook.

During some observations, coders could not observe information needed to determine whether to apply a particular code (such as whether a teacher formed student small groups randomly or strategically or whether instructional materials were developed by the core curriculum developer or the teacher). Following each in-person or video observation, we conducted a post-observation interview with the observed teacher using a semi-structured interview protocol, described above. Most interviews were conducted by a research team member other than the coder. Before each interview, the coder met with the interviewer to highlight codes that required clarification during the interview. Following each interview, coders reviewed post-observation teacher interview data to resolve outstanding coding issues.

A gold standard coder double-coded 10 percent of the video observations on a biweekly basis to assess coder drift. When IRR did not meet or exceed a Cohen's kappa (Cohen, 1960) standard for reliability of 80 percent, the coder and lead coder resolved coding discrepancies by discussion.

Interpreting and reporting AIM scores

The initial AIM tool contained 13 domains and 86 codes (or items) (refer to Appendix B for a complete list). To summarize and visualize data collected by the tool, the pre-validation version of the tool aggregates codes (or items) into domain scores. Although AIM domain scores are an indicator of how teachers use their instructional time, they do not reflect the total number of times we observed a practice within a specific interval and should not be interpreted in minutes. Instead, domain scores are reported as a "percentage of class time," specifically representing the percentage of intervals during which a practice or behavior was observed at least once. Refer to Appendix C for domain and item-level descriptives.

III. Validating the AIM Tool

In this section, we detail our data sources, sample, data collection activities, methods, and results.

Sample

To test the tool's psychometric properties, we used classroom observation data collected in late winter of school year (SY) 2021–2022 and SY 2022–2023. We partnered with four large urban districts to pilot test the tool. From the full study sample of 39 middle schools, we purposively constructed a sample of 13 "deep dive" schools to ensure representation across *district*, *school*, *study curricula*, and *grade levels* (grades 6–8). In the deep dive schools, we conducted 85 classroom observations:

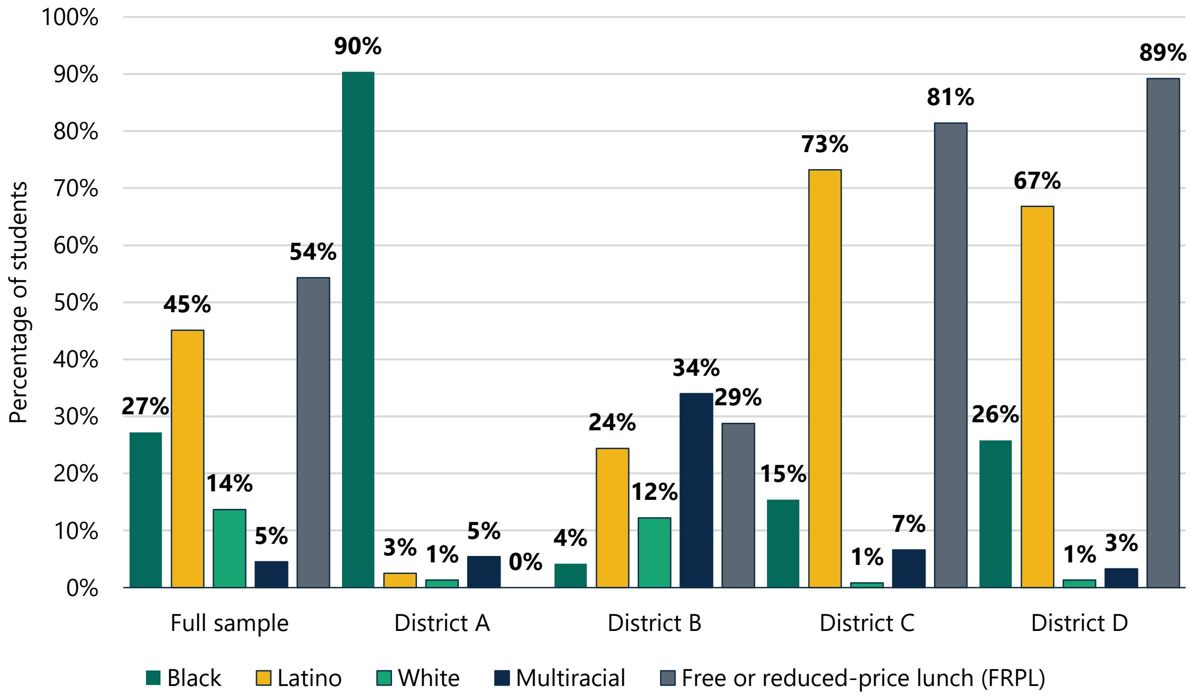
- 12 to 27 observations per **district** (4 districts; $\bar{x}$ = 21 observations)
- 1 to 11 observations per **school**[1] (13 schools; $\bar{x}$ = 6 observations)
- 2 to 39 observations per **curriculum** (6 curricula; $\bar{x}$ = 14 observations)
- 25 to 30 observations per **grade** (3 grades; $\bar{x}$ = 28 observations)
- 1 to 4 observations per **teacher** (39 teachers[2]; $\bar{x}$ = 2 observations)

In addition, we conducted 53 post-observation teacher interviews during which teachers reflected on the lesson observed to explain (1) the rationale behind instructional decisions and *adaptations* they made to the intended and planned curriculum, (2) whether and how *professional learning* activities influenced the observed lesson, and (3) their perspective on effective culturally responsive and equitable teaching. Refer to Exhibits III.1–III.2 for sample characteristics.

[1] One teacher from one school agreed to only one observation during the study's first year. Otherwise, we conducted at least three observations per school.

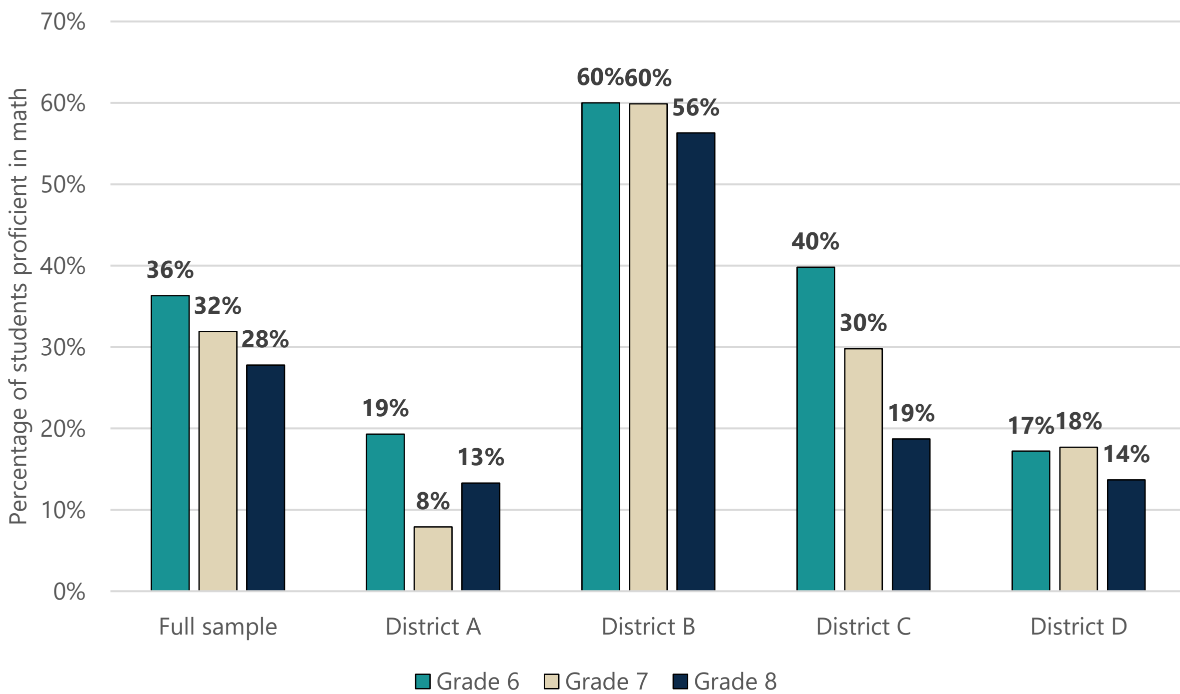
[2] The number of observations depended on whether a teacher taught more than one grade level, whether they participated in both data collection years, and their availability or willingness to be observed during the data collection window each school year.

Exhibit III.1. Student race/ethnicity and free and reduced-price lunch status



Source: SY 2020–2021 Common Core of Data. District A does not report FRPL data.

Exhibit III.2. Student math proficiency by grade



Source: SY 2020–2021 Common Core of Data

Exhibit III.3. Teacher characteristics

	Number	Percentage
Race/ethnicity (N = 56)		
Asian	2	3.6%
Black	2	3.6%
Hispanic	18	32.1%
Multiracial	3	5.4%
Native Hawaiian/Pacific Islander	2	3.6%
White	29	51.8%
Gender (N = 60)		
Prefer not to answer	2	3.3%
Female	33	55.0%
Male	25	41.7%
Highest degree earned (N = 60)		
Bachelors	19	31.7%
Masters	41	68.3%
Teaching credential (N = 62)		
Elementary teaching credential (multiple subjects)	20	32.3%
Math (middle school/grades 5–9) teaching credential in math	17	27.4%
Initial regular, standard, professional, or national board certification in my main teaching assignment	9	14.5%
Temporary, provisional, preliminary, probational, or emergency certification	7	11.3%
Math (secondary/grades 7–12) teaching credential in math	4	6.5%
Professional Elementary Teaching credential multi-subject 1–6, Special Education and Bilingual extension	3	4.8%
Regular, standard, or probationary certification not in my main teaching assignment	2	3.2%
Years teaching middle school math (N = 62)		
0–2 years	13	21.0%
3–5 years	13	21.0%
6–9 years	15	24.2%
Over 9 years	21	33.9%

Source: AMS Fall 2021 Teacher Survey.

Data collection

Classroom observation data

We were not permitted to video-record classrooms in one of the four study districts. Therefore, we used the AIM tool to conduct in-person observations in one school district and video observations in three districts. Between SY 2021–2022 and SY 2022–2023, we simplified and refined the tool's design and retrained all coders. At the conclusion of SY 2022–2023 data collection, we recoded all SY 2021–2022 video-recorded observations using the revised SY 2022–2023 version of the tool, resulting in data on 23 in-person observations in one school district and 62 video observations in the remaining districts. The final data set excludes SY 2021–2022 observation data collected in person because we cannot recode live observations.

Anticipating a need to validate our tool, we concurrently coded 76[3] of the classroom observations we conducted with the AIM tool in SY 2021–2022 and SY 2022–2023 using the Mathematics Scan observation tool (M-Scan; Bostic et al., 2021; Walkowiak et al., 2014, 2018). The M-Scan is a validated observation protocol designed to assess the degree to which teachers create opportunities for students to do the following:

1. Engage in cognitively demanding tasks
2. Identify, apply, and adapt a variety of strategies to solve problems
3. Connect mathematics to other mathematical concepts, their own experience, to the world around them, and to other disciplines
4. Use, contextualize, illustrate, and translate math ideas and concepts through multiple representations (such as pictures, graphs, symbols, and words)
5. Use mathematical tools (such as calculators, pattern blocks, fraction strips, counters, and virtual tools) to represent abstract mathematical concepts
6. Express their mathematical ideas openly and communicate their mathematical thinking clearly to their peers and teacher using the language of mathematics
7. Provide exp*lanations and justifications*, both orally and on written assignments

The M-Scan tool co-developers (the gold standard coders) trained three Mathematica coders—who were different from the AIM coders—on conducting observations with the tool in a five-day training. The gold standard coders had substantial classroom teaching experience and were math education experts. The three Mathematica coders had classroom teaching experience or had completed substantial coursework in math at the postsecondary level. The training involved reading, listening to conversations about each coding dimension, watching videos, and coding practice videos. The training involved a four-phase process: (1) preparation, (2) training and mastery, (3) reliability, and (4) drift test. The gold standard coders tracked and recorded trainees' progress in attaining and maintaining reliability through the four phases. Trainees practiced with the gold standard coders on at least two full class mathematics videos. After the training session, trainees watched two video-recorded classes independently and took notes. Afterward, trainees' ratings were compared to those of the gold standard coders. After trainees watched and coded the assigned set of "training" videos, the gold standard coders identified gaps and looked for convergence. More training videos were assigned if gaps were present. Trainees moved to the reliability phase when ratings from the training videos converged with ratings from the gold standard codes. Trainees watched and coded six mathematics "reliability" video observations, without conferring with the gold standard coder. After the gold standard coder verified that the trainee was reliable, the trainee was able to code mathematics observations using the M-Scan.

[3] We were unable to concurrently code nine of the in-person observations because those classes were taught in Spanish. Although one of our certified AIM coders is a fluent Spanish speaker, none of our certified M-Scan coders were fluent in Spanish.

To use the M-Scan tool, coders watch the first 30 minutes of a video-recorded lesson and take notes throughout the 30-minute segment to record what occurs during the lesson. Coders write their notes on the back of the coding sheet or on separate pieces of paper. The notes are used as examples and references when completing the M-Scan coding for that segment. After the first 30 minutes, the video is paused to allow coders to reflect and mark "soft codes" (that is, initial ratings) on the coding sheet by underlining the number corresponding to the initial code. These marks serve as indicators of what happened during the first part of the lesson. After assigning "soft codes" for the first 30 minutes, coders continue watching the lesson, following the procedures from the first 30-minute segment. Once coders have watched the entire lesson, they assign final codes of 1 to 7 to each dimension, where 1 and 2 represent a low rating (limited evidence of this domain), 3 to 5 represent a moderate rating, and 6 and 7 represent a high rating (more evidence and stronger in nature).

We used M-Scan's scoring rubrics to rate both the quality and frequency with which a teacher demonstrated each of the seven domains listed above during a lesson. We analyzed M-Scan ratings for a total of 76 lessons representing 25 different teachers; this included one to four observations per teacher, depending on whether they taught more than one grade level.

In-person observations were coded by one of the gold standard coders who co-developed the M-Scan and co-facilitated M-Scan training. All video observations were coded by one or two coders. During the data collection period, roughly 25 percent of the lessons were double-coded. In addition, the gold standard coders randomly checked for reliability on 20 percent of the videos, and they resolved coder discrepancies. To analyze M-Scan data, we calculated average ratings for each M-Scan domain.

Student survey

In the fall and spring of SY 2021–2022 and SY 2022–2023, we administered student surveys that asked students about their classroom experiences and beliefs, including their math enjoyment, engagement, self-efficacy, math achievement identity, and growth mindset. We analyzed survey data for 1835 students associated with the teachers we observed.

IV. Methods and Results

In this section, we detail our preliminary tests of the AIM tool's psychometric properties and present the results of these analyses. We collected evidence of *face validity* by expert review, assessed IRR using Cohen's kappa, evaluated the tool's *internal consistency* using Cronbach's alpha, and conducted two tests of validity: (1) *convergent and discriminant validity* using an existing observational measure of ambitious practice and (2) *construct validity* using student survey data to assess the extent to which our tool can measure inclusive practice. Lastly, we visualized the data produced by the tool to build evidence of the tool's capacity to measure ambitious and inclusive practice across instructional settings and curricula.

Face validity

Face validity refers to a type of validity based on a subjective judgement that a measure is covering the constructs that it aims to measure. In addition to conducting a literature review on research and evaluations of ambitious and inclusive practice (summarized above in the section on the AIM tool's design and development), we examined face validity by expert review. We consulted with members of our Math Advisory Council—comprising experts in mathematics education, professional learning, middle grades teaching and learning, and culturally responsive pedagogy—to vet and refine our list of initial codes, code descriptions, and code inclusion and exclusion criteria. These experts also reviewed and provided feedback on our initial approach to grouping codes into domains (such as student grouping strategies and performance tasks) and sub-domains (such as positive or negative relational interactions).

Interrater reliability

To ensure the tool can be consistently used to score a lesson across coders, teachers, lessons, curricula, and classrooms, we assessed interrater reliability using Cohen's kappa (Cohen, 1960). A senior member of the research team who led the development of the revised versions of the tool randomly selected 10 percent of all SY 2021–2022 observation data to conduct secondary, independent coding and resolved discrepancies with the initial coder by discussion. The same senior member of the team randomly double-coded 10 percent of all SY 2022–2023 observation data and all SY 2021–2022 video observations that were recoded using the SY 2022–2023 version of the tool to assess interrater reliability. Using Cohen's kappa, we estimated coder agreement as 89 percent at the domain level and 83 percent at the item level for the final data set.

Internal consistency of AIM teacher performance scales and AIM learning environment composite indicators

In this section, we discuss how we used AIM tool items and domains (groups of items) to construct and test the internal consistency of two types of measures:

1. **AIM teacher performance scales** to *evaluate*, classify, and compare the extent to which teachers use ambitious and inclusive practices
2. **AIM learning environment composite indicators** to *contextualize* the enabling and disabling conditions under which teachers use ambitious and inclusive practices

We designed these measures to simplify summarizing and reporting data collected with the AIM tool.

AIM teacher performance scales

Ultimately, we sought to develop reliable scales that we could use to evaluate, classify, and compare teacher performance with data collected with the AIM tool. To this end, we first explored the internal consistency of each AIM domain. Internal consistency estimates the extent to which a set of items that comprise a scale reliably measure the same construct. We used Cronbach's alpha (Cronbach, 1951), one of the most common methods for estimating internal consistency (Kimberlin & Winterstein, 2008). Cronbach's alpha considers the average intercorrelations of items and the number of items in a scale. Scales or domains constructed with a small number of items tend to perform poorly. Many of the initial AIM tool domains comprised just two items. For example, of the eight domains the tool defines as core ambitious and inclusive instructional practices, only three were found to be reliable due to low alphas. This indicates that the domains cannot be used reliably to assess teacher performance (Exhibit IV.1).

Exhibit IV.1. Reliability coefficients for the core AIM instructional practice domains

Core AIM instructional practice domains	Number of items	Cronbach's alpha
Real-world mathematical inquiry and problem solving	4	0.644
Multiple representations of mathematics	2	**0.829***
Mathematical discourse	4	**0.707***
Multilingual learner support and scaffolding	3	**0.819***
Engaged student and community funds of knowledge	2	0.052
Interdisciplinary connections	2	0.000
Empowered mathematical inquiry and decision making	2	NA

* Acceptable internal consistency ($\alpha > 0.70$).

NA = Domain has zero variance items.

Consequently, we iteratively constructed and tested the reliability of five performance scales that use items within and across multiple domains based on theorized relationships between AIM domains or items suggested by research on ambitious and inclusive teaching (Exhibit IV.2). We adjusted poorly performing scales (no or weak correlation between items) by discarding poorly correlated items ($r < 0.40$) to improve scale reliability (Exhibit IV.3). Appendix D details the items we used to construct each scale.

Exhibit IV.2. AIM teacher performance scale descriptions

Scale	Description
Ambitious practice	Cognitively demanding, standards-based instruction
Inclusive practice	*Culturally and linguistically responsive* (pedagogical knowledge, beliefs, dispositions, student expectations, and practices that collectively promote mathematical thinking, and the use of cultural and linguistic funds of knowledge as an instructional asset, and employ mathematics as a tool for social justice) and *equitable* (instructional protocols, tasks, or content that personalize or differentiates the learning experience for specific subgroups of students, such as multilingual learners, to ensure that all students have equal access and opportunity to engage in the learning process) instruction
Core AIM instructional practice	Teaching strategies promoted by the AIM tool that create opportunities for students to (1) engage in real-world mathematical inquiry and problem solving, (2) explore multiple representations of mathematics, (3) discuss mathematics in meaningful and rigorous ways, (4) develop academic literacy in mathematics as an English learner, (5) draw on their cultural and

Scale	Description
	community funds of knowledge as a learning asset, (6) make interdisciplinary connections, and (7) explore social justice issues of relevance to them using math as a tool
Student-centered practice	Classroom environments in which students participate in self-facilitated or self-directed math discussion, exploration, or performance tasks, often in peer pairs or small groups
Teacher-centered practice	Classroom environments in which a teacher is the primary focus of classroom interactions and lessons are largely delivered to the whole class with few opportunities for students to participate in self-directed learning in small groups or peer pairs

The scales we constructed—ambitious practice, inclusive practice, core AIM instructional practice, student-centered practice, and teacher-centered practice—were found to be reliable (α >0.70) based on a 0.70 to 0.95 range of acceptability (Tavakol & Dennick, 2011). Exhibit IV.3 presents reliability coefficients for the five performance scales.

Exhibit IV.3. Reliability coefficients for the AIM teacher performance scales

AIM teacher performance scales	Number of items	Cronbach's alpha
Ambitious practice	14	**0.819***
Inclusive practice	16	**0.756***
Core AIM instructional practice	17**	**0.816***
Student-centered practice	14	**0.773***
Teacher-centered practice	8	**0.733***

* Acceptable internal consistency (α >0.70).

** Excludes three zero variance items in the core AIM instructional practice domains: S_IC1, T_EM1, and S_EM1.
NA = Domain has zero variance items.

V. AIM learning environment composite indicators

To contextualize these performance scale scores, we developed composite indicators to characterize the learning environment in which teachers used (or did not use) ambitious and inclusive practices. We first grouped items that represent different characteristics or features of a learning environment into six sub-domains based on theorized or empirically supported relationships suggested by research on ambitious and inclusive instruction:

1. Positive relational interactions
2. Negative relational interactions
3. Administrative procedures and classroom protocols
4. Procedural instruction
5. High cognitive (performance tasks)
6. Low cognitive (performance tasks)

Refer to Appendix D for the list of items grouped into each sub-domain. As indicated in Exhibit IV.4, none of these sub-domains were found to be reliable (α < 0.70).

Exhibit IV.4. Reliability coefficients for sub-domains used to construct AIM learning environment composite indicators

AIM learning Environment sub-domains	Number of items	Cronbach's alpha
Positive relational interactions	13	0.592
Negative relational interactions	10	0.354
Administrative procedures and classroom protocols	5	0.449
Procedural instruction	3	0.224
High cognitive (performance tasks)	3	0.300
Low cognitive (performance tasks)	2	-0.31*

* Value is negative due to a negative average covariance among items violating reliability model assumptions. This suggests a need to either recode these items or reverse code them.

With these sub-domains, we constructed six composite indicators:

1. **Student grouping strategy gap:** When the gap is a positive value, students spent more time during an observed lesson working in peer pairs or small groups than in whole-class activities or doing independent desk work.
2. **Positive classroom culture:** When the value is positive, more positive than negative teacher–student and student–student interactions were observed.
3. **Ambitious–inclusive–procedural instruction ratio:** When the ratio is greater than 1, a teacher predominantly employed ambitious and/or inclusive practices during a lesson. When the ratio is less than 1, teachers predominantly employed procedural practices during a lesson.
4. **Student–teacher centeredness gap:** When the gap is a positive value, observed classroom practice was more student centered than teacher centered.
5. **High-low cognitive demand gap:** When the gap is a positive value, students participated in more high-cognitive demand than low-cognitive performance tasks during a lesson.
6. **Core-supplemental curriculum gap:** When the gap is positive, a teacher used the core curriculum more than supplemental materials during an observed lesson.

Appendix D outlines how we constructed, calculated, and interpreted each indicator.

Despite finding no evidence of reliability for the sub-domains used to construct these indicators, we believe they still hold promise and merit further investigation. For example:

- The **classroom culture** indicator is positively correlated with each of the five AIM scales for which we found evidence of reliability: *ambitious instructional practice* (r = 0.566; p = 0.000); *inclusive instructional practice* (r = 0.681; p = 0.000); *core AIM instructional practice* (r = 0.612; p = 0.000); *student-centered practice* (r = 0.676; p = 0.000); and *teacher-centered practice* (r = 0.485; p = 0.000). For instance, students in our sample experienced increasingly more positive classroom culture—more positive than negative interactions with their teacher and peers—the more their teacher used core AIM instructional practices (Exhibit IV.5).
- The **ambitious–inclusive–procedural instruction ratio** is positively correlated with school-level student math proficiency scores for *grade 6* (r = 0.423; p = 0.000), *grade 7* (r = 0.446; p = 0.000), and *grade 8* (r = 0.421; p = 0.000), as well as proficiency in English language arts for *grade 6* (r = 0.436; p =

0.000), *grade 7* (*r* = 0.426; p = 0.000), and *grade 8* (*r* = 0.421; p = 0.000). This suggests that the ratio could be an indicator of the coherence of the instructional climate in a school. For instance, the percentage of grade 6 students demonstrating proficiency in math is highest when the ratio is high—when teachers use ambitious and inclusive instructional practices more than procedural ones (Exhibit IV.6).

Exhibit IV.5. Classroom culture scores increased as the use of core AIM instructional practices increased

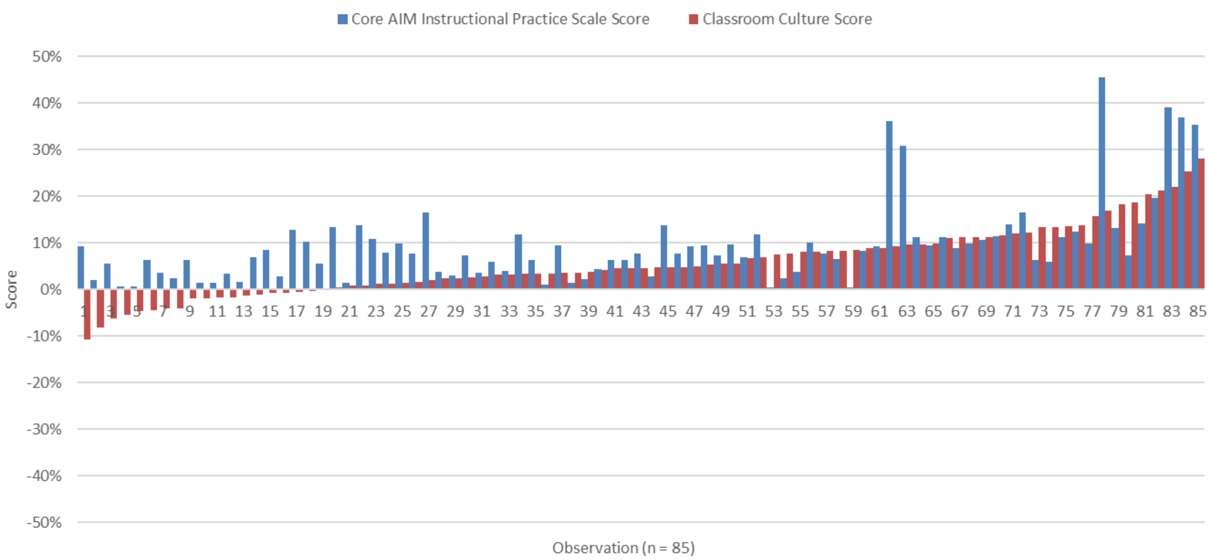
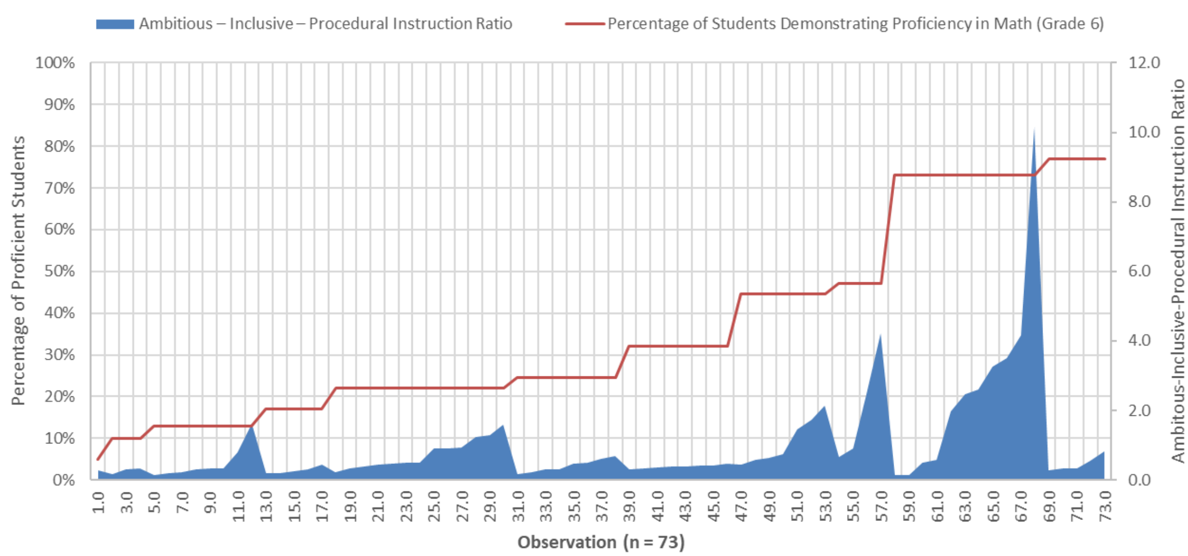


Exhibit IV.6. Ambitious–inclusive–procedural instruction ratios appear to be higher in schools in which there is a larger percentage of students demonstrating math proficiency



VI. Construct validity

We used student survey data to (1) assess the extent to which our tool can measure inclusive practice and (2) evaluate our hypothesis that procedural learning environments are less ambitious and can negatively influence student belief and engagement in math. Decades of research has demonstrated that students' beliefs about themselves and their mathematical abilities, as well as their enjoyment of mathematics, are strong predictors of mathematics performance (Exhibit IV.7).

Exhibit IV.7. Research on student beliefs about mathematics

Construct	Rationale and description
Beliefs • Growth mindset • Achievement identity • Math identity • Math enjoyment • Math self-efficacy	Students who value effort are said to have a **growth mindset**. The belief that our ability to learn is not fixed but can be developed over time is a mindset that can be nurtured in instructional settings (Burgoyne et al., 2018; Hochanadel & Finamore, 2015; Shanley et al., 2019; Yeager et al., 2019). In addition to a growth mindset, a number of other student beliefs can improve with intervention or are strong predictors of future mathematics achievement including students' (1) **identity** and self-concept as someone who can achieve academically (Lopez, 2017) and in math, **enjoyment** of mathematics (Goetz et al., 2008; Pinxten et al., 2014; Riegle-Crumb et al., 2011), and **self-efficacy** or confidence in solving mathematics problems and performing mathematics-related tasks. High self-efficacy is a predictor of mathematics achievement (Bandura, 1997; Evans, 2015; Hall & Ponton, 2005; Tarr et al., 2008; Warwick, 2008).
Engagement • Academic • Social/behavioral • Cognitive • Affective/motivational	Although low student engagement is commonly associated which negative academic outcomes such as dropping out of school, high student achievement is associated with positive **academic** (e.g., good grades and test scores), **social/behavioral** (e.g., high attendance), **cognitive** (e.g., improved conceptual understanding of a particular mathematics topic), and **affective** (e.g., intrinsic motivation to take rigorous coursework) outcomes (Finn & Zimmer, 2012; Fredricks & McColskey, 2012; Kong et al., 2003; Linnenbrink & Pintrich, 2003).

We constructed student survey scales for five constructs—math enjoyment, engagement, math achievement identity, math self-efficacy, and growth mindset—by calculating the average of all items associated with each construct. We calculated Cronbach's alpha to assess the reliability of each scale. First, we ensured that the alpha for each scale was equal to 0.70 or greater and that the alpha value would not be improved by removing any items. If either of these conditions was not met, we discussed as a group whether to remove any items from the scale. A list of the scales we created for the student surveys and the items that comprise each scale are listed in Appendix B.

We ran bivariate correlations between each of these scales and AIM tool scale scores, domain scores, subdomain scores, and items to test the AMS study's hypotheses that inclusive practice and positive relational interactions between students and teachers can foster students' math enjoyment, engagement, math achievement identity, math self-efficacy, and growth mindset. Comparatively, we tested the hypothesis that procedural learning environments are less ambitious and can negatively influence math enjoyment, engagement, math self-efficacy, math achievement identity, or growth mindset.

Influence of inclusive practice on students' classroom experiences

- Correlational analyses somewhat affirm our hypothesis that inclusive practice can foster students' math enjoyment, engagement, math achievement identity, math self-efficacy, and growth mindset. We found that AIM's inclusive practice performance scale is not associated with any of our non-academic student outcomes of interest. However, we also found the following:
- **The AIM inclusive instructional practice performance scale is *positively* associated with the AIM classroom culture composite indicator** (r = 0.768, p = 0.000) designed to measure supportive and inclusive relational interactions between students and teachers, such as those referenced in Exhibit IV.8.
- **A growth mindset is positively correlated with the core AIM instructional practice performance scale** (r = 0.574, p = 0.002), which includes such inclusive practices as *developing multilingual learners' academic literacy in mathematics* and *drawing on students' cultural and community funds of knowledge* as an asset for learning.

In addition, we found that our non-academic student outcomes of interest were positively correlated with culturally and linguistically responsive or equitable practices that are theorized to have a positive influence on student beliefs and experiences in math classrooms—for example, in Byrd's (2016) study of the relationship between student perceptions of their classroom experiences and culturally responsive pedagogy. (Exhibit IV.9).

Exhibit IV.8. Relationship between non-academic student outcomes and AIM measures of culturally responsive, linguistically responsive, and equitable instructional practice

Student outcome	Positively correlated student behaviors	Positively correlated teacher behaviors
Math enjoyment	• Requesting **English-language translation support** from a teacher (r = 0.466) or peer (r = 0.465)	• Setting a **positive emotional tone** (r = 0.529) • Giving **neutral feedback** (r = 0.408) • Using common, **non-technical language** (r = 0.507)
Engagement	• Participating in **small group activities** (r = 0.580)	• Engaging students' **cultural funds of knowledge** (r = 0.438) • Establishing or **reinforcing classroom norms** (r = 0.438) • Using scaffolding discourse (r = 0.415)
Math achievement identity	• Requesting **English-language translation support** from a teacher (r = 0.400) or peer (r = 0.405) • Requesting **assistance with math related tasks** from their teacher (r = 0.418) or peers (r = 0.472)	• Initiating **real-world inquiry** (r = 0.427) • Engaging students' **cultural funds of knowledge** (r = 0.454) • Giving **affirming feedback** (r = 0.506) • Giving **neutral feedback** (r = 0.511) • Establishing or **reinforcing classroom norms** (r = 0.437)

Student outcome	Positively correlated student behaviors	Positively correlated teacher behaviors
Math self-efficacy	• Exploring multiple representations (*r* = 0.413) • Requesting **English-language translation support** from a teacher (*r* = 0.442) or peer (*r* = 0.465) • Requesting **assistance with math-related tasks** from their teacher (*r* = 0.522) • Developing a **collective understanding** (*r* = 0.402, p = 0.038)	• Engaging students' **cultural funds of knowledge** (*r* = 0.478) • Giving **affirming feedback** (*r* = 0.412) • Giving **neutral feedback** (*r* = 0.605) • Establishing or **reinforcing classroom norms** (*r* = 0.511) • **Valuing math persistence** (*r* = 0.452, p = 0.018)
Growth mindset	• Developing a **collective understanding** (*r* = 0.425, p = 0.027)	• Engaging students' **cultural funds of knowledge** (*r* = 0.511) • Probing students to help them develop a **collective understanding** (*r* = 0.619, p = 0.001) • **Valuing math persistence** (*r* = 0.469, p = 0.014) • Making interpersonal connections (*r* = 0.535) • Establishing or **reinforcing classroom norms** (*r* = 0.570)

Notably, we also found that the AIM inclusive instructional practice performance scale is *positively* associated with the AIM ambitious instructional practice performance scale (*r* = 0.481, p = 0.000) that includes instructional practices such as engaging students in authentic problem solving, in mathematical discourse, and in tasks that require them to explain or justify their thinking. These results support our assertion that *inclusive and ambitious practice are complementary but not redundant. They measure distinct pedagogical strategies*. Moreover, the use of inclusive practices should not supplant a focus on rigor or learning standards.

We also found that **student growth mindset**—the belief that our ability to learn is not fixed but can be developed over time—is positively correlated with AIM's (1) ambitious instructional practice performance scale (*r* = 0.437, p = 0.23) and (2) ambitious-inclusive-procedural instruction ratio (*r* = 0.454, p = 0.017).

Influence of procedural instruction on students' classroom experiences

Correlational analyses did not affirm our hypothesis that procedural learning environments are less ambitious but do affirm our hypothesis that they can negatively influence math enjoyment, engagement, math achievement identity, or growth mindset. Although we found no evidence of a relationship between our procedural instruction measures and AIM's ambitious practice teacher performance scale or measures of low cognitive demand, we did find the following:

- **Initiation–response–evaluation (IRE) questioning**—the procedural practice that involves a teacher posing a question (for which there is a presumption of a "correct" or specific answer and that requires no elaboration or justification on the student's part) assesses the correctness of a student's response and gives close-ended feedback such as a yes or no—is *negatively correlated with math achievement identity* (*r* =-0.456). IRE is considered a low-cognitive form of mathematical discourse (Cazden, 1988;

Drageset, 2015; Park et al., 2020). We found that low-cognitive tasks that require students to memorize or recall math concepts or facts—such as IRE—are negatively associated with the following:

– Math achievement identity (*r* = -0.628)

– Math self-efficacy (*r* = -0.517)

– Math enjoyment (*r* = -0.683)

– Engagement (*r* = -0.652)

- **Lecturing or demonstrating**—a procedural practice that involves a teacher presenting, demonstrating, reviewing, defining, summarizing, or introducing instructional content in a non-interactive manner for an extended period of time—is *negatively correlated with math enjoyment* (*r* =-0.552) and *math achievement identity* (*r* =-0.554).

Convergent and discriminant validity

We used the M-Scan (Bostic et al., 2021; Walkowiak et al., 2014, 2018), an existing reliable and valid observational measure of ambitious practice that we reviewed at the outset of our study, to assess the convergent and discriminant validity of the AIM tool's ambitious and inclusive practice teacher performance scales. Recognizing that some M-Scan domains and AIM domains (domains used to construct the ambitious practice teacher performance scale) use similar terminology (such as "mathematical discourse" and "problem solving") but measure those constructs differently, we enlisted two M-Scan developers to review the AIM domains and sub-domains to identify M-Scan domains that we collectively theorize align with each other (convergent validity) as well as domains that we theorize measure different constructs (discriminant validity). We ran bivariate correlations between these domains.

As *evidence of convergent validity*, we theorized that six of the seven M-Scan domains would correlate with four of our AIM performance scales (Exhibit IV.9).

Exhibit IV.9. Theorized convergent M-Scan and AIM performance scales

M-Scan domains	AIM performance scales
Cognitive demand	• Ambitious instruction • Core AIM instructional practice
Problem solving	• Ambitious instruction • Core AIM instructional practice
Connections & applications	• Ambitious instruction • Core AIM instructional practice
Use of representations	• Ambitious instruction • Core AIM instructional practice
Use of math tools	• N/A (no theorized convergence)
Math discourse	• Ambitious instruction • Core AIM instructional practice • Teacher-centered practice • Student-centered practice

M-Scan domains	AIM performance scales
Explain & justify	• Ambitious instruction • Core AIM instructional practice • Teacher-centered practice • Student-centered practice

As *evidence of divergent validity*, we theorized that none of the M-Scan domains would correlate with the AIM inclusive practice scales because conceptually, they measure different types of pedagogical strategies. Using Dancey & Reidy's (2004) interpretation of Pearson's correlation coefficient for which 0.4 indicates a moderate association, we affirmed both our convergent and divergent validity assumptions (Exhibit IV.10).

Exhibit IV.10. Correlations between M-Scan domains and AIM performance scales

		AIM performance scales			
M-Scan domains	Ambitious instruction	Inclusive instruction	Core AIM instructional practice	Teacher-centered practice	Student-centered practice
Cognitive demand	**0.513***	0.202	**0.455***	**0.459***	0.348
Problem solving	**0.545***	0.237	**0.503***	**0.537***	0.373
Connections & applications	**0.430***	-0.064	0.380^	0.385	0.230
Use of representations	**0.444***	0.231	0.376^	0.376	0.303
Use of math tools	0.210	0.194	0.215	0.262	0.112
Math discourse	**0.533***	0.308	**0.496***	**0.448***	**0.418***
Explain & justify	**0.599***	0.258	**0.542***	**0.578***	**0.446***

Source: SY 2021–2022 and SY 2022–2023 classroom observation M-Scan and AIM domain scores (n = 76 video observations; excludes 10 in-person observations conducted in SY 2021–2022 that could not be recoded using the revised SY 2022–2023 version of the AIM tool).

Note: **Bold values** indicate scales for which convergence was anticipated and affirmed.

^ *V*alues indicated scales for which convergence was anticipated but not affirmed.

* $r > 0.40$; $p < .01$.

VII. Discussion

Results indicate that our tool is a promising measure of ambitious and inclusive instructional practice—cognitively demanding, standards-based mathematics instruction that is culturally responsive, linguistically responsive, and equitable. The results of our psychometric tests are summarized in Exhibit V.1.

Exhibit V.1. Summary of findings by research question

Research questions	Summary of findings
Is the AIM tool reliable and valid for use in a large-scale study across multiple contexts (curricula, districts, schools, classrooms, and instructional units)?	For a data set that spans four school districts, 39 classrooms, six curricula, and two consecutive school years: • We estimated **interrater reliability** as 89% at the domain level and 83% at the item level using Cohen's kappa. • We found the teacher **performance scales** we constructed to be reliable (α >0.70) (ambitious practice, inclusive practice, core AIM instructional practice, teacher-centered practice, and student-centered practice). • We found that three AIM performance scales that were designed to measure ambitious practice correlated with six M-Scan domains that were also designed to assess ambitious practice as **evidence of convergent validity**, • We found that our AIM inclusive practice performance scale did not correlate with any of the M-Scan domains as **evidence of divergent validity**, • **Our results suggest that inclusive and ambitious practice are complementary *but not redundant measures.*** *They evaluate distinct pedagogical strategies,* so the use of one should not supplant the other.
Is the AIM tool empirically supported? Do data collected by the AIM tool: • Support the assertion that inclusive practices are positively associated with ambitious practices? • Demonstrate that inclusive practice positively influences student belief and engagement in math? • Affirm our hypothesis that procedural learning environments are less ambitious and can negatively influence student belief and engagement in math?	• The AIM **ambitious practice performance scale** is positively correlated with AIM's inclusive practice performance scale. This result further supports our assertion that inclusive and ambitious practices are complementary. • AIM's **inclusive practice performance scale** is not positively associated with math enjoyment, engagement, math achievement identity, math self-efficacy, or growth mindset. However, we found that: • Each of the **non-academic student outcomes** (such as growth mindset and math achievement identity) we investigated is positively correlated with one or more AIM measures of culturally and linguistically responsive or equitable practices that are theorized to have a positive influence on student beliefs and experiences in math classrooms (such as engaging students' cultural funds of knowledge). • **Positive classroom culture** composite indicator scores appear to increase as the use of core AIM instructional practices increases. • AIM's measures of **procedural practice** do not correlate with AIM's measure of low-cognitive performance tasks but are negatively associated with math achievement identity, math self-efficacy, math enjoyment, and engagement.

Reflecting on these results, we updated our tool (refer to Appendix C for the revised codebook). A copy of the revised Excel-based tool is available for download at https://www.mathematica.org/-/media/B0CAB9E122F645619F40B4C0EC834757.ashx along with a version with sample data:

https://www.mathematica.org/-/media/47B47F0E9845472CBAB9BE116640E783.ashx.

Limitations

Six limitations influenced our ability to generate stronger evidence of the AIM classroom observation tool's reliability and validity:

- **Undersampling teachers who claimed to use inclusive practices.** Our full sample was constructed purposively based on whether a teacher taught one of the six study curricula. We excluded remedial and advanced course sections from consideration. From this purposive sample, and consistent with our partnership agreements with the participating districts, we only observed teachers who agreed to be observed and were available during the data collection window each school year. Unsurprisingly, we rarely observed some of the practices most commonly associated with culturally responsive teaching. This may have occurred for several reasons. First, although the teacher surveys we administered in the fall of each data collection year asked teachers to self-report their confidence with and frequency of using culturally responsive mathematics teaching strategies, COVID-related school closures in the year prior to conducting classroom observations limited our ability to collect and use teacher survey data in time to inform which teachers were selected for observation. As a result, we were unable to purposively construct a sub-sample of classrooms in which we might expect to observe culturally responsive practices. Second, we did not ask teachers to use culturally and linguistically responsive practices when we observed their classrooms. We stressed the importance of them teaching "business-as-usual." Third, teachers may define culturally responsive practice differently than does the AMS study. A future study might oversample teachers who claim to use inclusive practices frequently.
- **Conducting few observations of each teacher.** We only observed each teacher one to four times total during the two data collection years. The lessons we observed may not be a fair representation of a teacher's practice across the year or the frequency with which they use certain instructional practices. For instance, a teacher may be more likely to use culturally responsive practices at the end of an instructional unit, when students might complete a culminating formative assessment project that requires the use of the math concepts they learned during the unit to investigate a social justice issue of interest. Under ideal circumstances, research has recommended conducting at least three lesson observations by three different coders to gather formative data on teacher practice and at least 10 observations to make summative decisions about teacher performance (van der Lans et al., 2016).
- **Relying on student survey data to assess construct validity of AIM's inclusive practice measures.** Participating schools were responsible for administering student surveys for the full sample. Of the 186 teachers who responded to our surveys, schools administered surveys to just 46 percent of these teachers. Among the sample of teachers we observed, we were only able to obtain student survey data for 31 percent of the teachers. The incomplete data affect our ability to evaluate the validity of our inclusive practice performance scale and measures. In future studies, we need to strengthen our efforts to recruit student survey respondents or identify alternative approaches to testing those measures. We also considered that the student survey scales we used may need to be revisited. Although we constructed our scales from several *different* validated instruments, it's plausible that our scales are not sufficiently discriminant when administered together in the same survey. For example, the AMS study's non-academic outcomes of interest—math enjoyment, engagement, math achievement identity, self-

efficacy, and growth mindset—are not mutually exclusive concepts. For instance, *self-efficacy* and *growth mindset* both refer to an individual believing that they can effect positive change in their life through effort. We may revisit the student scale score construction in the future to confirm whether poor discriminant validity influenced our ability to build evidence of the construct validity of AIM's inclusive practice scale.

- **Assessing internal consistency with Cronbach's alpha.** Although the four scales we constructed—ambitious, inclusive, teacher-centered, and student-centered practice—were found to be reliable ($\alpha > 0.70$), only three of the AIM instructional practice domains met or exceeded the 0.7 threshold of acceptability for Cronbach's alpha. These less-than-ideal results were anticipated for several reasons. First, some of the items associated with the core AIM instructional domains were rarely or never observed, such as the empowered mathematical inquiry and decision-making domain. Cronbach's alpha cannot be calculated when there is no variance between items. Second, some of the domains are constructed with just two or three items. Cronbach's alpha is calculated by dividing the average covariance between items by the average total item variance. High alphas (or reliability) therefore require the covariance between items to be substantially higher than the item variance. As a result, domains with few items typically have lower alphas than domains with many items (Emons et al., 2007). Third, Cronbach's alpha is intended to measure the extent to which a set of items consistently measures the same concept. However, some of the AIM domains group items that measure different aspects of the same concept. For example, the performance tasks domain includes five items that represent five different types of student performance tasks that vary in increasing complexity from low cognitive demand to high. Finally, some researchers have argued that Cronbach's alpha is either inappropriately used or overused as an estimate of scale reliability—either overestimating the reliability of a scale or underestimating it results in rejecting a measure that may actually be reliable (Panayides, 2013; Taber, 2018; Zakariya, 2022).
- **Exploring the influence of the method of observation on the tool's reliability.** As discussed previously, we were not permitted to video-record classrooms in one of the four districts. Anecdotally, we did not experience a noteworthy difference in interrater reliability or audiovisual quality of the observation itself. However, we would like to examine in a future study whether and to what extent the method of observation influences the tool's reliability.
- **Investigating the influence of bias on AIM teacher performance scale scores across instructional contexts.** Factors such as the incoming academic performance of students, student course scheduling, and observer implicit biases about instructional norms and quality can influence the reliability of classroom observation scores (Bell et al., 2015; Campbell & Ronfeldt, 2018; Jones & Bergin, 2019; Liu et al., 2019; Luoto et al., 2023; Molina et al., 2018; Steinberg & Garrett, 2016). Although we explored the central tendency and variability of the AIM teacher performance scale scores (Appendix G), we did not evaluate whether and to what degree bias may influence the tool's ability to distinguish practice across instructional contexts (such as district, grade, teacher characteristics, student demographics, and curricula).

Future directions

With these limitations in mind, in a future study we would like to do the following:

1. **Test the AIM classroom observation tool with a larger, more diverse sample of teachers** that includes those who are committed to inclusive practice—perhaps coupled with an intervention that provides professional learning on the practices the AIM tool assesses. Classroom observation could be a tool to promote both ambitious and inclusive practice in middle school math classrooms. Although research on classroom observation as a professional learning intervention is a promising strategy (see for example, Cantrell et al., 2014), in a review by Bottiani et al. (2018) of the impact of in-service training on culturally responsive practice, there were not enough peer-reviewed studies employing causal research designs to make claims about their efficacy. To help address this research gap, we are in the early stages of adapting and piloting the AIM tool for use by teams of classroom teachers as an in-service professional learning intervention.
2. **Conduct more observations of each teacher over a longer period of time** (such as an entire instructional unit) and assess within teacher consistency.
3. **Conduct confirmatory factor and Rausch analyses** to reassess the AIM tool's internal consistency, further refine its structure and content, test the factor structure in different subgroups, and develop benchmarks for each teacher performance scale indicating low, moderate, and high performance levels.
4. Further explore the utility and validity of the AIM learning environment composite indicators.
5. **Revisit the construct validity of the inclusive practice measures,** including reassessing the reliability and validity of the student survey scales we constructed. Although we constructed our student survey scales from validated instruments, it is plausible that these scales are not sufficiently discriminant when used in the same survey.
6. **Conduct a sensitivity analysis** to identify biases that may influence AIM teacher performance scale scores across instructional contexts.
7. **Create a set of training materials and a certification program** to support high-fidelity use of the AIM tool. For the tool to be used reliably across instructional settings and curricula, we would like to develop a set of training materials, a coder training and certification program, and guidance on analyzing and interpreting AIM classroom observation data.
8. **Explore the potential of artificial intelligence to support classroom observation.** Collecting, coding, and analyzing observational data at scale—whether for research purposes or educator evaluation—is time intensive, resource intensive, and costly. In partnership with IRIS Connect, the platform we used to conduct video observations, we would also like to explore the potential of artificial intelligence (AI) to automate these tasks. However, whether AI can be used to reliably detect and assess ambitious and inclusive practice from video is an open question. The ability of AI tools to accurately interpret human behavior, interaction, and emotions is more emergent than established science. Researchers have demonstrated risks associated with using AI, such as algorithmic bias and facial, gender, and racial recognition discrimination. Some commercial applications of ML models have high error rates, and those errors have been found to disproportionately impact minoritized groups (Buolamwini & Gebru, 2018; Learned-Miller et al., 2020; Lee et al., 2019; Raji et al., 2022).

References

Aguirre, J. M., & del Rosario Zavala, M. (2013). Making culturally responsive mathematics teaching explicit: A lesson analysis tool. *Pedagogies: An International Journal*, *8*(2), 163–190.

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and instruction*, *16*(3), 183–198.

Aronson, B., & Laughter, J. (2016). The theory and practice of culturally relevant education: A synthesis of research across content areas. *Review of Educational Research*, *86*(1), 163–206.

Balfanz, R., Herzog, L., & MacIver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, 42(4), 223–235.

Bandura, A. (1997). Self-efficacy: The exercise of control. Freeman.

Battey, D., Leyva, L. A., Williams, I., Belizario, V. A., Greco, R., & Shah, R. (2018). Racial (mis) match in middle school mathematics classrooms: Relational interactions as a racialized mechanism. *Harvard Educational Review*, *88*(4), 455-482.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., Mccaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2015). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 50–97). John Wiley & Sons, Inc.

Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education, 24*(1), 5–31.

Boston, M., & Candela, A. G. (2018). The Instructional Quality Assessment as a tool for reflecting on instructional practice. *ZDM Mathematics Education, 50*, 427–444. https://doi.org/10.1007/s11858-018-0916-6

Boston, M., Bostic, J., Lesseig, K., & Sherman, M. (2015). A comparison of mathematics classroom observation protocols. *Mathematics Teacher Educator, 3*(2), 154–175. https://doi.org/10.5951/mathteaceduc.3.2.0154

Bottiani, J. H., Larson, K. E., Debnam, K. J., Bischoff, C. M., & Bradshaw, C. P. (2018). Promoting educators' use of culturally responsive practices: A systematic review of inservice interventions. *Journal of Teacher Education, 69*(4), 367–385.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research Conference on Fairness, Accountability and Transparency, 81,* 1-15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Burgoyne, A. P., Hambrick, D. Z., Moser, J. S., & Burt, S. A. (2018). Analysis of a mindset intervention. *Journal of Research in Personality, 77,*21–30.

Byrd, C. M. (2016). Does culturally relevant teaching work? An examination from student perspectives. *Sage Open*, *6*(3), 2158244016660744.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for. *American Educational Research Journal*, *55*(6), 1233–1267.

Cantrell, S. C., Correll, D. P., Malo-Juvera, V., & Ivanyuk, L. (2014). *Culturally responsive instruction observation protocol (CRIOP) professional development: Year 2*. Collaborative Center for Literacy Development.

Cazden, C. (1988). Classroom Discourse: The Language of Teaching and Learning. Porthmouth, NH: Heinemann.

Chen, P. P. (2003). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences*, *14*(1), 77–90.

Civil, M. (2016). STEM learning research through a funds of knowledge lens. *Cultural Studies of Science Education, 11*(1), 41–59.

Cleary, T. J., & Chen, P. P. (2009). Self-regulation, motivation, and math achievement in middle school: Variations across grade level and math context. *Journal of School Psychology*, *47*(5), 291–314.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Curenton, S. M., Iruka, I. U., Humphries, M., Jensen, B., Durden, T., Rochester, S. E., Sims, J., Whittaker, J. V., & Kinzie, M. B. (2019). Validity for the Assessing Classroom Sociocultural Equity Scale (ACSES) in early childhood classrooms. *Early Education and Development*, *31*(2), 284–303. https://doi.org/10.1080/10409289.2019.1611331

Dancey, Ch.P. & Reidy, J. (2004, 3rd Edition). Statistics without maths for Psychology: using SPSS for Windows. Pearson Education Limited, England.

de Araujo, Z., Roberts, S. A., Willey, C., & Zahner, W. (2018). English learners in K–12 mathematics education: A review of the literature. *Review of Educational Research, 88*(6), 879–919.

Desimone, L. M., & Garet, M. S. (2015). Best practices in teachers' professional development in the United States. Psychology, Society and Education, 7(3), 252–263.

Desimone, L. M., & Pak, K. (2017). Instructional coaching as high-quality professional development. Theory Into Practice, 56(1), 3–12.

Desimone, L., Hochberg, E. D., & McMaken, J. (2016). Teacher knowledge and instructional quality of beginning teachers: Growth and linkages. Teachers College Record, 118(5).

Drageset, O. G. (2015). Student and teacher interventions: A framework for analysing mathematical discourse in the classroom. Journal of Mathematics Teacher Education, 18, 253–272.

Edmonds-Wathen, C. (2019). Linguistic methodologies for investigating and representing multiple languages in mathematics education research. Research in Mathematics Education, 21(2), 119–134.

Emons, W. H., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. Psychological Methods, 12(1), 105.

Erath, K., Ingram, J., Moschkovich, J., & Prediger, S. (2021). Designing and enacting instruction that enhances language for mathematics learning: A review of the state of development and research. ZDM Mathematics Education, 53(2), 245–62.

Evans, J. A. (2015). Gender, self-efficacy, and mathematics achievement: An analysis of fourth grade and eighth grade TIMSS data from the United States. Educational Studies Dissertations,63. https://digitalcommons.lesley.edu/education_dissertations/63

Finn, J. D., & Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? In Handbook of research on student engagement (pp. 97–131). Springer.

Franco, M. P., Bottiani, J. H., & Bradshaw, C. P. (2024). Assessing teachers' culturally responsive classroom practice in PK–12 schools: A systematic review of teacher-, student-, and observer-report measures. Review of Educational Research, 94(5), 743–798.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In Handbook of research on student engagement (pp. 763–782). Springer.

Goetz, T., Frenzel, A. C., Hall, N. C., & Pekrun, R. (2008). Antecedents of academic emotions: Testing the internal/external frame of reference model for academic enjoyment. Contemporary Educational Psychology, 33(1), 9-33.

Hall, J. M., & Ponton, M. K. (2005). Mathematics self-efficacy of college freshman. Journal of Developmental Education, 28(3), 26.

Hanzlian, C. G. (2013). Using a modified cultural relevance rubric to assess and implement culturally relevant texts in content area classrooms for ELLs [Master's thesis, State University of New York at Fredonia].

Hochanadel, A., & Finamore, D. (2015). Fixed and growth mindset in education and how grit helps students persist in the face of adversity. Journal of International Education Research, 11(1), 47–50.

Jackson, K., & Cobb, P. (2010, April). Refining a vision of ambitious mathematics instruction to address issues of equity [Presentation]. Annual Meeting of the American Educational Research Association, Denver, CO, United States.

Jitendra, A. K., Griffin, C. C., Haria, P., Leh, J., Adams, A., & Kaduvettoor, A. (2007). A comparison of single and multiple strategy instruction on third-grade students' mathematical problem solving. Journal of Educational Psychology, 99(1), 115.

Jones, E., & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. Educational Assessment, 24(2), 91–118.

Jones, S. (2015). Mathematics teachers' use of the culturally relevant cognitively demanding mathematics task framework and rubric in the classroom. Northeastern Educational Research Association Conference Proceedings, 12.

Kazemi, E., Franke, M., & Lampert, M. (2009, July). Developing pedagogies in teacher education to support novice teachers' ability to enact ambitious instruction. Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia, 1,12–30.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. American Journal of Health-System Pharmacy, 65(23), 2276–2284.

Kong, Q. P., Wong, N. Y., & Lam, C. C. (2003). Student engagement in mathematics: Development of instrument and validation of construct. Mathematics Education Research Journal, 15(1), 4–21.

Larios, R. J., Karras, J. E., Suárez-Orozco, C., & Bashir-Baaqee, I. S. J. (2022). Using an iterative approach to systematically observe culturally responsive practices across classrooms. Urban Education, 60(1). https://doi.org/10.1177/00420859221139832

Learned-Miller, E., Ordóñez, V., Morgenstern, J., & Buolamwini, J. (2020). Facial recognition technologies in the wild. Algorithmic Justice League. https://people.cs.umass.edu/~elm/papers/FRTintheWild.pdf

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. Journal of Mathematics Teacher Education, 14, 25–47. https://doi.org/10.1007/s10857-010-9140-1

Lee, N., Resnick, P., & Barton, G. (2019). Algorithmic bias detection and mitigation: Best clinical and policies to lower consumer harms. The Brookings Institution. https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Linnenbrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs in student engagement and learning in the classroom. Reading &Writing Quarterly, 19(2), 119–137.

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. Educational Assessment, Evaluation and Accountability, 31, 61–95.

Lopez, F. A. (2017). Altering the trajectory of the self-fulfilling prophecy: Asset-based pedagogy and classroom dynamics. Journal of Teacher Education, 68(2), 193–212.

Luoto, J., Klette, K., & Blikstad-Balas, M. (2023). Possible biases in observation systems when applied across contexts: Conceptualizing, operationalizing, and sequencing instructional quality. Educational Assessment, Evaluation and Accountability, 35(1), 105–128.

Marshall, J., Smart, J., & Horton, R. (2010). The design and validation of EQUIP: An instrument to assess inquiry-based instruction. International Journal of Science and Mathematics Education, 8, 299–321. https://link.springer.com/article/10.1007/s10763-009-9174-y

Mason, L., & Scrivani, L. (2004). Enhancing students' mathematical beliefs: An intervention study. *Learning and Instruction*, *14*(2), 153–176.

Molina, E., Fatima, S. F., Ho, A. D. Y., Melo Hurtado, C. E., Wilichowksi, T., & Pushparatnam, A. (2018). Measuring teaching practices at scale: Results from the development and validation of the TEACH classroom observation tool (Policy research working paper no. WPS 8653). World Bank Group. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/464361543244734516/measuring-teaching-practices-at-scale-results-from-the-development-and-validation-of-the-teach-classroom-observation-tool Moschkovich, J. (2013). Equitable practices in mathematics classrooms: Research-based recommendations. Teaching for Excellence and Equity in Mathematics, 5(1).

Moschkovich, J. N. (2015). Academic literacy in mathematics for English learners. The Journal of Mathematical Behavior, 40, 43–62.

National Academies of Sciences, Engineering, and Medicine. (2018). English learners in STEM subjects: Transforming classrooms, schools, and lives. The National Academies Press. https://doi.org/10.17226/25182

Nowicki, J. M. (2022). Pandemic learning: As students struggled to learn, teachers reported few strategies as particularly helpful to mitigate learning loss. Report to Congressional Committees (GAO-22-104487). US Government Accountability Office.

Panayides, P. (2013). Coefficient alpha: Interpret with caution. Europe's Journal of Psychology, 9(4).

Pape, S. J., & Tchoshanov, M. A. (2001). The role of representation(s) in developing mathematical understanding. Theory into Practice, 40(2), 118–127.

Park, M., Yi, M., Flores, R., & Nguyen, B. (2020). Informal formative assessment conversations in mathematics: Focusing on preservice teachers' initiation, response and follow-up Sequences in the classroom. Eurasia Journal of Mathematics, Science and Technology Education, 16(10).

Pinxten, M., Marsh, H. W., De Fraine, B., Van Den Noortgate, W., & Van Damme, J. (2014). Enjoying mathematics or feeling competent in mathematics? Reciprocal effects on mathematics achievement and perceived math effort expenditure. British Journal of Educational Psychology, 84(1), 152-174.

Powell, R., Cantrell, S. C., Malo-Juvera, V., & Correll, P. (2016). Operationalizing culturally responsive instruction: Preliminary findings of CRIOP research. Teachers College Record, 118(1), 1–46. https://doi.org/10.1177/016146811611800107

Powell, R., Cantrell, S. C., Rightmyer, E. (2013). Teaching and reaching all students: An instructional model for closing the gap. Middle School Journal. May 2013. https://www.wsra.org/assets/Rebecca%20Powells%20Article%20in%20the%20Middle%20School%20Journal.pdf

Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022, June). The fallacy of AI functionality. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 959–972. https://dl.acm.org/doi/pdf/10.1145/3531146.3533158

Riegle-Crumb, C., & Grodsky, E. (2010). Racial-ethnic differences at the intersection of math course-taking and achievement. *Sociology of Education*, *83*(3), 248–270.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R. & Bloom, I. (2002), Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. School Science and Mathematics, 102, 245 –253. https://doi.org/10.1111/j.1949-8594.2002.tb17883.x

Schommer-Aikins, M., Duell, O. K., & Hutter, R. (2005). Epistemological beliefs, mathematical problem-solving beliefs, and academic performance of middle school students. *The Elementary School Journal*, *105*(3), 289–304.

Sciarra, D. T. (2010). Predictive factors in intensive math course-taking in high school. *Professional School Counseling*, *13*(3), 2156759X1001300307.

Selling, S. K. (2016). Making mathematical practices explicit in urban middle and high school mathematics classrooms. Journal for Research in Mathematics Education, 47(5), 505-551.

Shanley, L., Biancarosa, G., Clarke, B., & Goode, J. (2019). Relations between mathematics achievement growth and the development of mathematics self-concept in elementary and middle grades. Contemporary Educational Psychology, 59, 101804. https://doi.org/10.1016/j.cedpsych.2019.101804

Sleeter, C. (2012). Confronting the marginalization of culturally responsive pedagogy. Urban Education, 47, 562–584. https://doi.org/10.1177/0042085911431472

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure. Educational Evaluation and Policy Analysis, 38(2), 293–317.

Stroupe, D. (2016). Beginning teachers' use of resources to enact and learn from ambitious instruction. Cognition and Instruction, 34(1), 51–77.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in science education, 48, 1273–1296.

Tarr, J. E., Reys, R. E., Reys, B. J., Chávez, Ó., Shih, J., & Osterlind, S. J. (2008). The impact of middle-grades mathematics curricula and the classroom learning environment on student achievement. Journal for Research in Mathematics Education, 39(3), 247–280.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53.

Turner, E. E., Drake, C., McDuffie, A. R., Aguirre, J., Bartell, T. G., & Foote, M. Q. (2012). Promoting equity in mathematics teacher preparation: A framework for advancing teacher learning of children's multiple mathematics knowledge bases. Journal of Mathematics Teacher Education, 15(1), 67–82.

van der Lans, R. M., van de Grift, W. J., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. Studies in Educational Evaluation, 50, 88–95.

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. Educational Studies in Mathematics, 85(1), 109–128.

Walkowiak, T. A., Berry, R. Q., Pinter, H. H., & Jacobson, E. D. (2018). Utilizing the M-Scan to measure standards-based mathematics teaching practices: Affordances and limitations. ZDM Mathematics Education, 50(3), 461–474.

Wang, J., & Goldschmidt, P. (2003). Importance of middle school mathematics on high school students' mathematics achievement. *The Journal of Educational Research*, *97*(1), 3–17.

Warwick, J. (2008). Mathematical self-efficacy and student engagement in the mathematics classroom. MSOR Connections, 8(3), 31–37.

Womack, Sue Ann. (2011). Measuring mathematics instruction in elementary classrooms: Comprehensive Mathematics Instruction (CMI) observation protocol development and validation (Publication no. 2905) (Theses and Dissertations, Brigham Young University). https://scholarsarchive.byu.edu/etd/2905

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., ... Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. Nature, 573(7774), 364–369.

Zakariya, Y. F. (2022). Cronbach's alpha in mathematics education research: Its appropriateness, overuse, and alternatives in estimating scale reliability. Frontiers in Psychology, 13, 1074430.

Appendix A:
Existing observation methods

This page has been left blank for double-sided copying.

Reform-Oriented Teaching Observation Protocol (RTOP)

The RTOP (Sawada et al., 2002; Boston et al., 2015) was developed to support education reform efforts in professional development and teacher education and was used by the Arizona Collaborative for Excellence in the Preparation of Teachers (ACEPT). It was designed to measure the degree to which mathematics and science teaching are *reform oriented*, which they define as standards-based teaching, an inquiry orientation in lesson design and implementation, and student-centered teaching practices. The tool is a 25-item questionnaire using five-point Likert scales for each item and examining Lesson Design and Implementation, Content, and Classroom Culture (communicative interactions and student–teacher relationships). We determined it was not useful for our purposes because it does not have items or domains specifically focused on culturally responsive practices.

Instructional Quality Assessment (IQA) in Mathematics

The IQA (Boston & Candela, 2018; Boston et al., 2015) was designed to measure the quality of mathematics instruction at scale using a combination of lesson observations, assignment collections, and student work. The IQA is based on two main constructs: *Academic Rigor* and *Accountable Talk*; the IQA assesses the quality of instruction based on the mathematical work that students *do* and *discuss* in the classroom, based on the cognitive demands and accountable talk moves observed during the lesson. There are multiple rubrics describing specific practices within each major construct (for example, Academic Rigor has rubrics for *Potential of the Task, Task Implementation, and Rigor of the Discussion)*. Each rubric is scaled from 0 to 4. We determined it was not useful for our purposes because it does not have items or domains specifically focused on culturally responsive practices.

Mathematical Quality of Instruction (MQI)

The MQI (Learning Mathematics for Teaching Project, 2011; Boston et al., 2015) is a multidimensional assessment of the rigor and richness of the mathematics present during classroom instruction. It was developed alongside efforts to conceptualize and validate measures of mathematical knowledge for teaching. The instrument is organized around five dimensions of instruction: *Classroom Work is Connected to Mathematics*, *Richness of the Mathematics*, *Working with Students and Mathematics, Errors and Imprecision*, and *Common Core Aligned Student Practices.* There are subscales within each dimension. Videotaped lessons are divided into equal intervals of 5 to 7.5 minutes, with each segment coded yes/no or on a scale of 0 to 3 along the five dimensions. We determined it was not useful for our purposes because it does not have items or domains specifically focused on culturally responsive practices.

Comprehensive Mathematics Instruction (CMI) Observation Protocol

The CMI (Womack, 2011) uses embedded teaching and learning cycles to build students' mathematical understanding using a guided inquiry approach. The steps in the CMI framework are (1) develop understanding, (2) solidify understanding, and (3) practice understanding. The authors developed an observation protocol aligned to the framework in partnership with a CMI expert panel, pilot tested the protocol in 12 classrooms, and then validated the protocol with a larger sample of 144 classrooms. All items are scored on a five-point rating scale, with quality evaluations. The protocol is made up of six sections. Three are aligned to the standard sections of a CMI lesson: launch, explore, and discuss. Three could happen at any point in the lesson: mathematics content, classroom climate, and lesson coherence.

We determined it was not useful for our purposes because it does not have items or domains specifically focused on culturally responsive practices and because it is scored inferentially, rather than deductively.

Electronic Quality of Inquiry Protocol (EQUIP)

The EQUIP (Marshall et al., 2010) was designed to measure the quantity and quality of inquiry-based instruction. It is composed of 26 indicators measured within three constructs: instruction (for example, conceptual development, order of instruction), curriculum (for example, content depth, assessment type), and ecology (for example, classroom discourse, visual environment). We determined it was not useful for our purposes because it does not have items or domains specifically focused on culturally responsive practices and because it is scored inferentially, rather than deductively.

Mathematics Scan (M-Scan)

The M-Scan (Walkowiak et al., 2014) is a validated observation protocol designed to assess the degree to which teachers create opportunities for students to engage in cognitively demanding tasks; identify, apply, and adapt a variety of strategies to solve problems; connect mathematics to other mathematical concepts, to their own experience, to the world around them, and to other disciplines; use, contextualize, illustrate, and translate math ideas and concepts through multiple representations (such as pictures, graphs, symbols, and words); use mathematical tools (such as calculators, pattern blocks, fraction strips, counters, and virtual tools) to represent abstract mathematical concepts; express their mathematical ideas openly and communicate their mathematical thinking clearly to their peers and teacher using the language of mathematics; and provide explanations and justifications, both orally and on written assignments. In addition, the M-Scan assesses the degree to which teachers structure a lesson to be conceptually coherent, so that activities are connected mathematically and build on one another in a logical manner as well as present mathematical concepts and model mathematical discourse clearly and accurately throughout the lesson. The tool is scored on a scale from 1 to 7, with values further summarized as low (1–2), medium (3–5), and high (6–7). Detailed rubric descriptions correspond to numerical ratings. We determined it was not useful for our purposes because it does not have items or domains specifically focused on culturally responsive practices and because it is scored inferentially, rather than deductively.

We used the M-Scan to assess the AIM tool’s convergent and discriminant validity.

Assessing Classroom Sociocultural Equity Scale (ACSES)

The ACSES (Curenton et al., 2019) is composed of five major factors: Challenging Status Quo Knowledge, Equitable Learning Opportunities for racially minoritized learners, Equitable Discipline, Connections to Home Life, and Personalized Learning Opportunities. Scoring for each dimension is based on the frequency of occurrence and how many students it affected: 1 (never) = Did not exhibit; 2 (hardly ever) = Exhibited 1 time or with only a few children; 3 (sometimes) = Exhibited 2–3 times with some children; 4 (very often) = Exhibited often with about half children, but inconsistently; 5 (nearly always) = Exhibited consistently with nearly all children. Higher scores indicated more equitable learning opportunities after the necessary items are reverse scored. We determined it was not useful for our purposes because it was developed and tested only in early childhood classrooms (Pre-K to grade 3).

Systematic approach to culturally responsive practices (CRP) across classrooms

The CRP tool (Larios et al., 2022) builds on qualitative review of teacher practices to identify and holistically assess culturally responsive practices that can be systematically observed across multiple classrooms. This tool was developed and refined over several rounds of data collection to ensure practices are empirically supported. Observers score the entire lesson on a scale of –2 to +2, where –2 = Actively culturally hostile; –1 = deficit lens; 0 = absence of CRP; +1 = contributive approach; and +2 = Additive approach. We determined it was not useful for our purposes because it is not math specific, and because it is scored inferentially rather than deductively.

Culturally Responsive Instruction Observation Protocol (CRIOP)

The CRIOP (Powell et al., 2016; Powell et al., 2013) describes and measures culturally responsive instruction using seven key domains: Classroom Relationships, Family Collaboration, Assessment, Curriculum/Planned Experiences, Instruction/Pedagogy, Discourse/Instructional Conversation, and Sociopolitical Consciousness/Diverse Perspectives. Assessment of classroom practice is measured using a four-point scale: 1 = not at all; 2 = occasionally; 3 = often; and 4 = to a great extent. This tool was implemented in the context of a professional learning program, in which participating teachers received coaching, on-site professional development, and support with instructional planning focused on a CRIOP framework. We determined it was not useful for our purposes because it is not math specific or middle school specific.

Appendix B:
Initial AIM tool codes and descriptions

This page has been left blank for double-sided copying.

Real-world mathematical inquiry and problem solving (RWMI)		
T_RWMI1	Facilitate real world inquiry	TEACHER poses a mathematical question, problem, or task with explicit real-world implications or that requires applying real-world data or information to solve.
S_RWMI1	Initiate real world inquiry	STUDENT(S) poses a mathematical question/problem or task with explicit real-world implications or that requires applying real-world data or information to solve.
S_RWMI2D	Discuss real world problem or data	STUDENT(S) discuss a mathematical question/problem, data, or information with explicit real-world implications or that requires applying real-world data or information to solve.
S_RWMI2P	Participate in real world inquiry	STUDENT(S) participate in a math task with explicit real-world implications or that requires applying real-world data or information to solve.

Multiple representations of mathematics (MRM)		
T_MRM1	Model multiple representations	TEACHER thinks out loud to demonstrate the kinds of questions students should ask themselves to reason or make sense of different symbolic, textual, or graphical representations of mathematical concepts or relationships OR to share their rationale or justification for different solution paths.
T_MRM2	Explore multiple representations	TEACHER probes, asks purposeful questions, or provides instructions for a math task that encourages students to share, discuss or demonstrate (1) their reasoning and sense making about different symbolic, textual, or graphical representations of mathematical concepts or relationships, (2) connections or relationships of the mathematical concepts, procedures, or tasks at hand with other mathematical ideas (e.g., presented in a different lesson), or (3) alternative solution paths.
S_MRM2	Explore multiple representations	STUDENTS share, discuss, or demonstrate (1) their reasoning and sense making about different symbolic, textual, or graphical representations of mathematical concepts or relationships, (2) connections or relationships of the mathematical concepts, procedures, or tasks at hand with other mathematical ideas (e.g., presented in a different lesson), or (3) alternative solution paths with other students.

Mathematical discourse (MD)		
T_MD1	Model the use of math terminology	TEACHER explicitly models, reviews, or prompts students to use math terminology, typically terms that are specific to the observed lesson or instructional unit.
S_MD1	Use of math terminology	STUDENT(S) use math terminology, typically terms that are relevant to the observed lesson or instructional unit.
T_MD2	Use of common, non-technical language	TEACHER uses non-math-specific vocabulary or verbal shorthand to discuss mathematical concepts or procedures.
S_MD2	Use of common, non-technical language	STUDENT(S) uses non-math-specific vocabulary to discuss mathematical concepts or procedures.
T_MD2	Developing a collective understanding	TEACHER probes, asks purposeful questions, or provides instructions to engage more than one student to (1) evaluate or compare each other's representations, solutions, approaches, or arguments, (2) debate math ideas and strategies, or (3) co-construct strategies or explanations in response to a mathematical task.

Mathematical discourse (MD)		
S_MD2	Developing a collective understanding	More than one STUDENT (in large, small, or peer pair groups) (1) evaluate or compare each other's representations, solutions, approaches, or arguments, (2) debate math ideas and strategies, or (3) co-construct strategies and explanations in response to a mathematical task.

Multilingual learner support and scaffolding (ELSS)		
T_ELSS1	Use of English language scaffolding strategies	TEACHER uses an English-language scaffolding strategy or provides linguistic support to make a math-related conversation or task more accessible. This code does not presume English as the predominate language; it refers to supporting students' linguistic understanding and fluency regardless of language.
S_ELSS2	Requests translation support	MULTILINGUAL STUDENT asks a TEACHER for language support, such as what an English word or math term means or how to say something in English.
S_ELSS3	Peer language support	STUDENT(S) asks or offers translation support to another student, or students engage in on-task conversation in a language other than English.

Engaged student and community funds of knowledge (FoK)		
T_FoK1	Cultural funds of knowledge	TEACHER connects or employs students' community, cultural, or linguistic knowledge that is specific to their individual lived experience or local context with a math-related discussion or task.
S_FoK1	Cultural funds of knowledge	STUDENT(S) connects or employs community, cultural, or linguistic knowledge that is specific to their individual lived experience or local context with a math-related discussion or task.

Interdisciplinary connections (IC)		
T_IC1	Make interdisciplinary connection	TEACHER explicitly connects a math-related discussion or task to another academic discipline or content area (e.g., science, social studies, art) as a tool to broaden students' understanding and application of a mathematical fact, concept, or procedure beyond the lesson.
S_IC1	Make interdisciplinary connection	STUDENT(S) connects a math-related discussion or task to another academic discipline or content area (e.g., science, social studies, art) as a tool to broaden students' understanding and application of a mathematical fact, concept, or procedure beyond the lesson.

Empowered mathematical inquiry and decision making (EMI)		
T_EMI1	Facilitate empowered mathematical inquiry	TEACHER poses a question, initiates a discussion, or assigns an instructional task that requires students to use math to investigate or critique a societal challenge or a social justice issue of direct relevance to them or of their own choosing.
S_EMI1	Engage in empowered mathematical inquiry	STUDENT(S) use math to investigate or critique a societal challenge or a social justice issue of direct relevance to them or of their own choosing.

Relational interactions (RI)		
T_RI1P	Addressing student behavior	TEACHER praises student(s)' *positive* non-math-related or on-task behavior.
T_RI1N	Addressing student behavior	TEACHER redirects or reprimands student(s)' *negative, noncompliant, or off-task* non-math student behavior.
T_RI2P	Framing mathematics ability	TEACHER makes a comment that *positively* frames one or more students' general capabilities in mathematics or ability to complete an upcoming math task. Instances must include broad statements rather than a specific assessment of a contribution during the lesson.
T_RI2N	Framing mathematics ability	TEACHER makes a comment that *negatively* frames one or more students' general capabilities in mathematics or ability to complete an upcoming math task. Instances must include broad statements rather than a specific assessment of a contribution during the lesson.
S_RI2P	Framing mathematics ability	STUDENT makes a comment that *positively* frames their own or another student's general capabilities in mathematics or ability to complete an upcoming math task.
S_RI2N	Framing mathematics ability	STUDENT makes a comment that *negatively* frames their own or another student's general capabilities in mathematics or ability to complete an upcoming math task.
T_RI4P	Setting the emotional tone	TEACHER sets *positive* expectations for the classroom culture/climate by preempting behavioral issues with compassion and empathy or creating a safe emotional space for students.
T_RI4N	Setting the emotional tone	TEACHER sets *negative* expectations for the classroom culture/climate by preempting behavioral issues with threats, warnings, or other statements of *negative* consequences.
T_RI5	Scaffolding discourse	TEACHER provides math-related feedback, asks questions, or models the thinking process to help a student break down a cognitively demanding or complex task into more manageable, accessible, or comprehensible parts.
S_RI6T	Requesting assistance	STUDENT asks a teacher for math-related help with a lesson-related activity that advances their understanding of a math concept or ability to complete a mathematical procedure.
S_RI6S	Requesting assistance	STUDENT asks *another student* for math-related help with a lesson-related activity that advances their understanding of a math concept or ability to complete a mathematical procedure.
T_RI7P	Valuing math persistence and a growth mindset	TEACHER encourages students to work through cognitively demanding tasks by *praising* confusion and mistakes or encouraging productive struggle.
T_RI7N	Devaluing math persistence	TEACHER discourages working through cognitively demanding tasks by *reprimanding or ridiculing* struggle, confusion, and mistakes.
T_RI7D	Discomfort with productive struggle	TEACHER demonstrates discomfort with one or more students' struggling to complete an instructional task by jumping in to help shortly after assigning a task.
S_RI7P	Valuing math persistence	STUDENT expresses about themselves or encourages others to work through cognitively demanding tasks by *praising* struggle, confusion, and mistakes.

Relational interactions (RI)		
S_RI7N	Devaluing math persistence	STUDENT expresses about themselves or discourages others from working through cognitively demanding tasks by *reprimanding or ridiculing* struggle, confusion, and mistakes.
T_RI8M	Correcting	Correcting a student's *math-related* misconceptions, error, or misstep by sharing the correct answer or demonstrating the appropriate approach.
T_RI8NM	Correcting	Correcting *non-math-related* errors (e.g., grammar, pronunciation, vocabulary)
T_RI9	Moderating the amount of speech	TEACHER urges student(s) to speak less or more when discussing math-related ideas or content.
T_RI10	Rhetorical questioning	TEACHER asks a rhetorical math-related question for which they do not expect a response.
T_RI11	Non-inclusive instructional decision	TEACHER makes an instructional decision that could be perceived as unrelatable, problematic, or inappropriate by one or more students in the classroom.
T_RI12P	Giving Affirming Feedback	Teacher gives a student positive, supportive, or constructive feedback on their math-related work or contributions—but does not elaborate or explore as to why the work is good.
T_RI12N	Giving Negative Feedback	Teacher gives a student negative, unconstructive, or unsupportive feedback on their math-related work or contributions—but does not explain why the work is poor.
T_RI12NT	Giving Neutral Feedback	Teacher gives a student feedback that does not evaluate, confirm, or refute the accuracy of their answer. The teacher simply acknowledges that the student has offered a response or made a contribution.
T_RI13	Interpersonal connection	TEACHER forges or reinforces a personal or relational connection with one or more students via a shared interest, expressing curiosity or appreciation for a student’s interest, or engaging with a student in their home language.

Procedural practice (PP)		
T_PP1	Taking attendance	TEACHER verbally or nonverbally takes attendance, counts students, or otherwise indicates that they are taking note of present/absent students.
T_PP2	Collecting homework/classwork	TEACHER physically or digitally collects student work.
T_PP3	Assigning homework/classwork	TEACHER assigns a homework or classwork assignment.
T_PP4	Making an announcement	TEACHER makes an announcement that is not related to the current math lesson.
T_PP5	Establishing or reinforcing classroom norms	TEACHER explains, discusses, or reminds students of classroom procedures, rules, or code of conduct that is not specific to the math lesson.
T_PP6	Initiation-Response-Evaluation (IRE) questioning	TEACHER poses a question—for which there is a presumption of a "correct" or specific answer and that requires no elaboration or justification on the student's part—assesses the correctness of a student's response, and gives close-ended feedback such as a yes/no.

Procedural practice (PP)		
T_PP7	Lecturing or demonstrating	TEACHER presents, demonstrates, reviews, defines, summarizes, or introduces instructional content in a non-interactive manner for an extended period of time.
T_PP8	Procedural clarification	TEACHER provides a clarification or reminder about the instructions for a lesson-related activity that has already been assigned.
S_PP8	Procedural clarification	STUDENT(S) asks the teacher a procedural, non-math-related question to clarify expectations for an activity.
S_PP9	Warm up/close out	STUDENT(S) complete a brief and procedurally normed activity at the opening or closing of the class.

Performance tasks (PT)		
S_PT1	Memorize or recall	STUDENT(S) commit to memory or reproduce previously learned facts, rules, formulas, or definitions without connection to the concepts or meaning that underlie.
S_PT2	Perform procedures	STUDENT(S) use an algorithm or procedure to solve a problem with a focus on producing correct answers. No explanation is required, or explanations focus solely on describing the procedure that was used.
S_PT3	Demonstrate understanding	STUDENT(S) focus on the use of procedures for the purpose of developing understanding of mathematical concepts and ideas or providing explanations for why steps in a procedure make sense.
S_PT4	Conjecture, generalize, or prove	STUDENT(S) notice patterns or make observations and use these to form a conclusion; they engage in complex, non-algorithmic thinking to explore and understand the nature of mathematical concepts, processes, or relationships.
S_PT5	Solve non-routine problems or making connections	STUDENT(S) use relevant knowledge and experiences to work through a novel task or a task that could be represented or solved in multiple ways; student makes connections among various representations or strategies.

Grouping (G)		
G1	Whole class	TEACHER facilitates an instructional task, discussion, or presentation to the entire class.
G2	Small group	TEACHER assigns an instructional task to one or more students to be completed in small groups (groups of 3-8 students) based on proximity, classroom norms, or student choice OR assigns students to groups to personalize or differentiate the math learning environment, such as by learning need, learning preference, or ability OR assigns students to different learning stations, typically distinguishable by different learning activities and locations within the classroom.
G3	Pair	TEACHER assigns an instructional task to one or more students to be completed in pairs (groups of 2 students).
G4	Individual	TEACHER assigns an instructional task to one or more students to be completed independently (in isolation from or without support or collaborating with other students).

Instructional materials (IM)		
IM1	Textbook/Workbook	TEACHER or STUDENTS interact with a textbook or associated workbook manufactured by a curriculum company.
IM2P	Worksheet/handout: Paper-based	STUDENTS interact with a structured paper-based document with instructions, tasks, and space for students to complete work.
IM2E	Worksheet/handout: Electronic	STUDENTS interact with a structured electronic document with instructions, tasks, and space for students to complete work.
IM3	Blackboard/whiteboard/smartboard/overhead	TEACHER or STUDENTS interact with a large board or screen, visible to all students in the room, to facilitate whole-class learning. This may include a document camera or transparency machine.
IM4	Audio-visual recording	TEACHER or STUDENTS interact with a video clip, audio clip, or digital timer.
IM5I	Computer/Tablet: Individual	STUDENTS interact with individual or personal devices to access curriculum content, complete activities, or submit work.
IM6P	Assessment: Paper-based	STUDENTS interact with paper-based assessments or tools for teachers to gauge student learning.
IM6C	Assessment: Electronic	STUDENTS interact with electronic assessments or tools for teachers to gauge student learning.
IM7	Learning management system or other educational technology	TEACHER or STUDENTS interact with *or reference* a digital learning management system or another educational technology tool.
IM8A	Manipulative: Analog	STUDENTS interact with physical objects which support learning or engagement with a specific math concept.
IM8D	Manipulative: Digital	STUDENTS interact with digital tools which support learning or engagement with a specific math concept.
IM9	Unstructured materials	STUDENTS interact with blank or unstructured materials with no scaffolding or written structure.
IM10	Other	TEACHER or STUDENTS interact with or reference any other instructional material. Note what it is in the running records.

Instructional material type (IMT)		
IMT1	Core curriculum	The primary textbook the teacher is instructed to use by the school or district. The core curriculum should be one of the 6 study curriculums.
IMT2SD	Supplemental: Curriculum/learning platform developer	Content or materials developed by a curriculum or learning platform developer that is not part of the core curriculum. This may include purchased or free materials.
IMT2SS	Supplemental: State or district developed	Content or materials developed by the teacher's state or district that are not part of the core curriculum. This may include pacing charts or guidance about standards to prioritize.
IMT2ST	Supplemental: Teacher developed	Content or materials developed by the teacher observed or by another teacher.
IMT2SO	Supplemental: Other	Any other content or materials developed by a source not captured in an above code.
IMT3C	Culturally Responsive	Content or material that incorporates culturally responsive content or is a culturally responsive artifact—whether or not it furthers math learning.

Instructional material type (IMT)		
IMT3L	Language aid for multilingual learner	Content or material that has been adapted to support multilingual learners.

Appendix C:
AIM domain and item-level descriptives

This page has been left blank for double-sided copying.

	N	Min	Max	Mean	Std. deviation
Total intervals	85	7	18	12.3	3.2
Total observed class time (minutes)	85	40	90	61.5	15.6
Intervals using core curriculum	85	0	18	6.8	5.9
Intervals using supplemental materials	85	0	17	6.6	4.7
Real-world mathematical inquiry and problem solving (RWMI)	85	0.0%	65.6%	9.4%	13.8%
T_RWMI1	85	0.0%	100.0%	16.8%	24.2%
S_RWMI1	85	0.0%	28.6%	1.0%	4.4%
S_RWMI2D	85	0.0%	87.5%	5.1%	15.4%
S_RWMI2P	85	0.0%	100.0%	14.6%	27.0%
Multiple representations of mathematics (MRM)	85	0.0%	70.8%	10.2%	15.3%
MRM_Teacher	85	0.0%	56.3%	10.6%	13.6%
MRM_Student	85	0.0%	100.0%	9.4%	24.3%
T_MRM1	85	0.0%	50.0%	9.8%	13.8%
T_MRM2	85	0.0%	87.5%	11.5%	20.3%
S_MRM2	85	0.0%	100.0%	9.4%	24.3%
Mathematical discourse (MD)	85	0.0%	75.0%	18.8%	16.0%
T_MD1	85	0.0%	100.0%	30.9%	24.4%
S_MD1	85	0.0%	90.0%	17.3%	21.5%
T_MD3	85	0.0%	100.0%	19.5%	23.2%
S_MD3	85	0.0%	100.0%	7.3%	18.1%
Multilingual learner support and scaffolding (ELSS)	85	0.0%	66.7%	4.7%	12.1%
T_ELSS1	85	0.0%	100.0%	10.6%	25.2%
S_ELSS2	85	0.0%	75.0%	2.1%	10.4%
S_ELSS3	85	0.0%	57.1%	1.5%	7.6%
Engaged student and community funds of knowledge (FoK)	85	0.0%	11.1%	0.5%	1.9%
T_FoK1	85	0.0%	16.7%	0.7%	2.7%
S_FoK1	85	0.0%	22.2%	0.3%	2.5%
Interdisciplinary connections (IC)	85	0.0%	50.0%	1.7%	7.9%
T_IC1	85	0.0%	100.0%	3.4%	15.8%
S_IC1	85	0.0%	0.0%	0.0%	0.0%
Empowered mathematical inquiry and decision-making (EMI)	85	0.0%	0.0%	0.0%	0.0%
T_EMI1	85	0.0%	0.0%	0.0%	0.0%
S_EMI1	85	0.0%	0.0%	0.0%	0.0%
Relational interactions (RI)	85	1.2%	21.9%	9.5%	4.1%
Sub-Domain: RI_Positive	85	2.6%	31.7%	13.3%	6.7%
Sub-Domain: RI_Negative	85	0.0%	20.0%	7.6%	4.5%

	N	Min	Max	Mean	Std. deviation
T_RI1P	85	0.0%	61.5%	8.2%	13.8%
T_RI1N	85	0.0%	100.0%	39.5%	29.4%
T_RI2P	85	0.0%	36.4%	4.4%	8.2%
T_RI2N	85	0.0%	18.2%	1.0%	3.2%
S_RI2P	85	0.0%	9.1%	0.2%	1.3%
S_RI2N	85	0.0%	11.1%	0.6%	2.3%
T_RI4P	85	0.0%	44.4%	6.9%	10.7%
T_RI4N	85	0.0%	38.9%	2.5%	6.3%
T_RI5	85	0.0%	100.0%	44.9%	27.2%
S_RI6T	85	0.0%	100.0%	23.7%	22.0%
S_RI6S	85	0.0%	50.0%	3.3%	8.6%
T_RI7P	85	0.0%	38.5%	9.0%	11.5%
T_RI7N	85	0.0%	22.2%	1.3%	4.2%
T_RI7D	85	0.0%	36.4%	4.2%	8.7%
S_RI7P	85	0.0%	12.5%	0.4%	2.1%
S_RI7N	85	0.0%	11.1%	0.5%	2.2%
T_RI8M	85	0.0%	77.8%	19.7%	16.6%
T_RI8NM	85	0.0%	18.2%	2.0%	4.4%
T_RI9	85	0.0%	33.3%	4.2%	7.8%
T_RI10	85	0.0%	60.0%	10.8%	17.1%
T_RI11	85	0.0%	16.7%	0.4%	2.3%
T_RI12P	85	0.0%	100.0%	33.3%	25.4%
T_RI12N	85	0.0%	33.3%	2.6%	6.3%
T_RI12NT	85	0.0%	87.5%	7.1%	15.7%
T_RI13	85	0.0%	72.7%	6.7%	12.4%
T_RIMD2	85	0.0%	91.7%	11.5%	18.2%
S_RIMD2	85	0.0%	66.7%	6.1%	11.6%
Procedural practice (PP)	85	4.5%	35.6%	17.7%	6.7%
Sub-Domain: Administrative Procedures	85	4.5%	38.9%	16.8%	7.0%
Sub-Domain: Procedural Instruction	85	2.0%	47.2%	21.2%	11.7%
T_PP1	85	0.0%	12.5%	3.7%	4.4%
T_PP2	85	0.0%	44.4%	6.3%	11.0%
T_PP3	85	0.0%	88.9%	30.4%	18.2%
T_PP4	85	0.0%	33.3%	5.2%	6.7%
T_PP5	85	0.0%	70.0%	18.1%	16.9%
T_PP6	85	0.0%	100.0%	41.3%	25.1%
T_PP7	85	0.0%	66.7%	6.0%	12.3%
T_PP8	85	0.0%	100.0%	37.1%	23.8%
S_PP8	85	0.0%	60.0%	12.8%	14.9%
S_PP9	85	0.0%	66.7%	16.1%	16.1%

	N	Min	Max	Mean	Std. deviation
Performance tasks (PT)	85	3.6%	52.7%	23.5%	8.1%
Sub-Domain: HighCognitive	85	0.0%	58.8%	13.8%	14.1%
Sub-Domain: LowCognitive	85	0.0%	100.0%	38.2%	20.3%
S_PT1	85	0.0%	100.0%	18.4%	22.8%
S_PT2	85	0.0%	100.0%	58.0%	34.0%
S_PT3	85	0.0%	100.0%	31.8%	32.4%
S_PT4	85	0.0%	81.8%	7.7%	17.5%
S_PT5	85	0.0%	50.0%	1.7%	8.7%
Grouping (G)	85	9.1%	45.0%	23.9%	6.4%
Sub-Domain: Teacher Directed (G1 and G4)	85	0.0%	87.5%	51.4%	17.6%
Sub-Domain: Student Directed (G2 and G3)	85	0.0%	42.5%	10.2%	9.3%
G1_Whole Class	85	0.0%	100.0%	61.8%	24.4%
G2_SmallGroup	85	0.0%	56.7%	9.8%	11.6%
G3_Peer Pair	85	0.0%	81.8%	11.4%	20.4%
G4_Individual	85	0.0%	100.0%	41.0%	28.7%
Instructional materials (IM)	84	8.5%	33.0%	17.9%	5.7%
IM1	85	0.0%	100.0%	16.1%	30.2%
IM2P	85	0.0%	100.0%	35.6%	35.9%
IM2E	85	0.0%	83.3%	1.4%	9.7%
IM3	85	0.0%	100.0%	72.2%	26.7%
IM4	85	0.0%	71.4%	2.9%	10.5%
IM5I	85	0.0%	100.0%	21.5%	36.6%
IM6P	84	0.0%	100.0%	4.5%	13.5%
IM6C	85	0.0%	22.2%	0.5%	3.2%
IM7	85	0.0%	100.0%	14.0%	32.2%
IM8A	85	0.0%	90.0%	11.9%	22.6%
IM8D	85	0.0%	100.0%	13.8%	28.1%
IM9	85	0.0%	100.0%	35.7%	37.0%
IM10	85	0.0%	75.0%	2.5%	10.5%
Instructional material type (IMT)	85	0.0%	42.9%	17.4%	7.2%
IMT1	85	0.0%	100.0%	54.3%	42.7%
IMT2SD	85	0.0%	100.0%	9.5%	25.4%
IMT2SS	85	0.0%	40.0%	0.5%	4.4%
IMT2ST	85	0.0%	100.0%	45.4%	43.4%
IMT2SO	85	0.0%	100.0%	8.2%	21.3%
IMT3C	85	0.0%	0.0%	0.0%	0.0%
IMT3L	85	0.0%	100.0%	3.8%	18.6%

This page has been left blank for double-sided copying.

Appendix D:
AIM teacher performance scale and AIM learning environment composite indicator construction and composite indicator descriptives

This page has been left blank for double-sided copying.

Exhibit D.1. Final AIM performance scale construction

Scale	Definition	Associated items
Ambitious practice	Cognitively demanding, standards-based instruction	• S_MD1: Student use of math terminology • S_MD3: Student Developing a collective understanding • S_MRM2: Student Explore multiple representations • S_PT3: Student Demonstrate understanding • S_PT4: Student Conjecture, generalize, or prove • S_PT5: Student Solve non-routine problems or making connections • S_RWMI1: Student Initiate real world inquiry • S_RWMI2D: Student Discuss real world problem or data • S_RWMI2P: Student Participate in real world inquiry • T_MD1: Teacher Model the use of math terminology • T_MD3: Teacher Developing a collective understanding • T_MRM1: Teacher Model multiple representations • T_MRM2: Teacher Explore multiple representations • T_RWMI1: Teacher Facilitate real world inquiry
Inclusive practice	*Culturally and linguistically responsive* (pedagogical knowledge, beliefs, dispositions, student expectations, and practices that collectively promote mathematical thinking, use cultural and linguistic funds of knowledge as an instructional asset, and employ mathematics as a tool for social justice) and *equitable* (instructional protocols, tasks, or content that personalizes or differentiates the learning experience for specific subgroups of students, such as multilingual learners, to ensure that all students have equal access and opportunity to engage in the learning process) instruction	• G2: Small Group • G3: Pairs • IM8A: Analog Manipulative • S_ELSS2: Student Requests translation support • S_ELSS3: Student Offers translation support • S_MD1: Student Use of math terminology • S_MD2: Student Use of common, non-technical language • S_MD3: Student Developing a collective understanding • S_MRM2: Student Explore multiple representations • S_PP9: Student Warm up/close out • S_RWMI1: Student Initiate real world inquiry • S_RWMI2D: Student Discuss real world problem or data • S_RWMI2P: Student Participate in real world inquiry • T_ELSS1: Teacher Use of English language scaffolding strategies • T_MD2: Teacher Use of common, non-technical language • T_RI5: Teacher Scaffolding discourse

Scale	Definition	Associated items
Core AIM	Teaching strategies that create opportunities for students to (1) engage in real-world mathematical inquiry and problem solving, (2) explore multiple representations of mathematics, (3) discuss mathematics in meaningful and rigorous ways, (4) develop academic literacy in mathematics as an English learner, (5) draw on their cultural and community funds of knowledge as a learning asset, (6) make interdisciplinary connections, and (7) explore social justice issues of relevance to them using math as a tool	S_ELSS2: Student Requests translation support S_ELSS3: Student Offers translation support S_FoK1: Student Cultural funds of knowledge S_MD1: Student Use of math terminology S_MD3: Student Developing a collective understanding S_MRM2: Student Explore multiple representations S_RWMI1: Student Initiate real world inquiry S_RWMI2D: Student Discuss real world problem or data S_RWMI2P: Student Participate in real world inquiry T_ELSS1: Teacher Use of English language scaffolding strategies T_FoK1: Teacher Cultural funds of knowledge T_IC1: Teacher Make interdisciplinary connection T_MD1: Teacher Model the use of math terminology T_MD3: Teacher Developing a collective understanding T_MRM1: Teacher Model multiple representations T_MRM2: Teacher Explore multiple representations T_RWMI1: Teacher Facilitate real world inquiry
Student-centered practice	Classroom environments in which students participate in self-facilitated or self-directed math discussion, exploration, or performance tasks, often in peer pairs or small groups	G2: Small Group S_ELSS2: Student Requests translation support S_ELSS3: Student Offers translation support S_MD1: Student Use of math terminology S_MD2: Student Use of common, non-technical language S_MD3: Student Developing a collective understanding S_MRM2: Student Explore multiple representations S_PP9: Student Warm up/close out S_RWMI1: Student Initiate real world inquiry S_RWMI2D: Student Discuss real world problem or data S_RWMI2P: Student Participate in real world inquiry T_ELSS1: Teacher Use of English language scaffolding strategies T_MD2: Teacher Use of common, non-technical language T_RI5: Teacher Scaffolding discourse
Teacher-centered practice	Classroom environments in which a teacher is the primary focus of classroom interactions and lessons are largely delivered to the whole class, with few opportunities for students to participate in self-directed learning in small groups or peer pairs	G1: Whole class S_PT3: Student Demonstrate understanding S_RWMI2D: Student Discuss real world problem or data T_MD1: Teacher Model the use of math terminology T_MD3: Teacher Developing a collective understanding T_MRM1: Teacher Model multiple representations T_MRM2: Teacher Explore multiple representations T_RWMI1: Teacher Facilitate real world inquiry

Exhibit D.3. AIM learning environment sub-domains

AIM learning environment sub-domains	Associated codes
Positive relational interactions	• T_RI2P: (Positively) framing mathematics ability • S_RI2P: (Positively) framing mathematics ability • T_RI4P: Setting a positive emotional tone • T_RI5: Scaffolding discourse • S_RI6T: Requesting assistance • T_RI7P: Valuing math persistence and a growth mindset • T_RI12P: Giving Affirming Feedback • T_RI12NT: Giving Neutral Feedback • T_RI13: Interpersonal connection • T_RIMD2: Use of common, non-technical language • S_RIMD2: Use of common, non-technical language • T_PP5: Establishing or reinforcing classroom norms
Negative relational interactions	• T_RI1N: Addressing student behavior • T_RI2N: (Negatively) framing mathematics ability • S_RI2N: (Negatively) framing mathematics ability • T_RI4N: Setting a negative emotional tone • T_RI7N: Devaluing math persistence and a growth mindset • T_RI7D: Discomfort with productive struggle • S_RI7N: Devaluing math persistence • T_RI8M: Correcting (math related) • T_RINM: Correcting (non-math related) • T_RI9: Moderating the amount of speech
Administrative procedures and classroom protocols	• T_PP1: Taking attendance • T_PP2: Collecting homework/classwork • T_PP3: Assigning homework/classwork • T_PP4: Making an announcement • T_PP8: Procedural clarification
Procedural instruction	• T_PP6: Initiation-Response-Evaluation (IRE) questioning • T_PP7: Lecturing or demonstrating • S_PP9: Warm up/close out
High cognitive (performance tasks)	• S_PT3: Demonstrate understanding • S_PT4: Conjecture, generalize, or prove • S_PT5: Solve non-routine problems or making connections
Low cognitive (performance tasks)	• S_PT1: Memorize or recall • S_PT2: Perform procedures

Exhibit D.5. AIM learning environment composite indicator calculations and interpretation

Domain	Associated variables	Calculation	Interpretation
Student grouping strategy gap	• Whole class • Small group • Peer pair • Individual	[Percentage of intervals during which students participate in peer pairs or small group activities] – [percentage of intervals during which students participate in whole class activities or independent desk work]	Positive value indicates students spent more time during an observed lesson working in peer pairs or small groups than in whole-class activities or doing independent desk work
Classroom culture	• Positive relational interactions sub-domain • Negative relational interactions sub-domain	[Positive relational interactions sub-domain score] – [negative relational interactions sub-domain score]	Positive value indicates students experienced more positive interactions with their teacher and peers than negative interactions during an observed lesson; negative value indicates students experience more negative than positive interactions
Ambitious–inclusive–procedural instruction ratio	• Ambitious instruction scale • Inclusive instruction scale • Procedural instruction sub-domain	[(ambitious instructional scale score + inclusive instructional scale score) ÷ 2] ÷ [procedural instructional sub-domain score]	Ratios greater than 1 indicate practice that is dominated by ambitious and/or inclusive practice; ratios less than 1 indicate practice that is dominated by procedural instruction; ratios that equal 1 indicate practice that balances ambitious, inclusive, and procedural instruction.
Student–teacher centeredness gap	• Student-centered practice scale • Teacher-centered practice scale	[Student-centered practice scale score] – [teacher-centered practice scale score]	Positive value indicates classroom practice was more student-centered than teacher-centered during an observed lesson
High-low cognitive demand gap	• High cognitive demand sub-domain • Low cognitive demand sub-domain	[High cognitive demand subdomain score] – [low cognitive demand subdomain score]	Positive value indicates students participated in high cognitive demand tasks more than low cognitive demand activities during an observed lesson

Domain	Associated variables	Calculation	Interpretation
Core-supplemental curriculum gap	• Core curriculum • Supplemental: – Curriculum/learning platform developer – State or district developed – Teacher developed – Other	[Percentage of intervals during which a teacher uses the core curriculum] – [percentage of intervals during which a teacher uses supplemental materials]	Positive value indicates teacher used the core curriculum more than supplemental materials during an observed lesson; zero value indicates teacher used the core curriculum as much as supplemental materials

Exhibit D.6. AIM teacher performance scale score and AIM learning environment composite indicator descriptives

	N	Min	Max	Mean	Std. Deviation
AIM teacher performance scale score descriptives					
Ambitious practice	85	0.0%	62.5%	13.2%	11.4%
Inclusive practice	85	1.5%	42.7%	11.3%	8.7%
Core AIM instructional practice	85	0.4%	45.6%	9.5%	8.8%
Student-centered practice	85	0.6%	46.1%	11.2%	9.3%
Teacher-centered practice	85	2.8%	78.1%	23.4%	13.5%
AIM learning environment composite indicator descriptives					
Student grouping strategy gap	85	-87.5%	25.0%	-41.2%	23.3%
Classroom culture	85	-10.7%	28.0%	5.7%	7.4%
Ambitious–inclusive-procedural instruction ratio	85	5.3%	1014.7%	93.1%	134.6%
Student-teacher centeredness gap	85	-32.2%	9.6%	-12.2%	10.0%
High-low cognitive demand gap	85	-100.0%	58.8%	-24.5%	30.2%
Core-supplemental curriculum gap	85	-17	16	0.2	8.7

Appendix E:
Student survey scales

This page has been left blank for double-sided copying.

Construct	Survey	Construct definition	Survey items included*
Student engagement scale	Fall and spring student surveys	Positive and active participation in math class, including the desire to meet academic expectations (such as earning good grades and test scores), comply with social and behavioral classroom norms (such as being a good small group partner), engage cognitively (such as the personal drive or commitment to improve conceptual understanding of a particular math topic), and engage emotionally (such as being excited when playing math games)	When reading the following statements, think about your current math class and decide how well the statements describe you. a. I don't think that hard when I am doing work for math class. b. I complete my math homework on time. c. I don't participate in math class. d. I do other things when I am supposed to be paying attention. e. I try to work with others who can help me in math. f. I build on others' ideas. g. I try to understand other people's ideas in math class. h. I don't care about other people's ideas. When reading the following statements, think about your current math class and decide how well the statements describe you. a. I try to understand my mistakes when I get something wrong. a. I want to understand what is learned in math class. b. I try to help others who are struggling in math. c. I talk about math outside of class. d. I think that math class is boring. e. I don't like working with classmates.
Math enjoyment scale	Fall and spring student surveys	The belief that doing math and being in math class is fun	When reading the following statements, think about your current math class and decide how well the statements describe you. a. I look forward to math class. b. I enjoy learning new things about math. c. I feel good when I am in math class. d. I often feel frustrated in math class. e. I don't care about learning math. f. I don't want to be in math class. g. I often feel down when I am in math class. h. I get worried when I learn new things about math.

Construct	Survey	Construct definition	Survey items included*
Math self-efficacy scale	Fall and spring student surveys	Students' confidence in solving math problems and performing math-related tasks; high self-efficacy is a predictor of math achievement	When reading the following statements, think about your current math class and decide how well the statements describe you. How much do you disagree or agree with the statements below? 2. I learn things quickly in math. 3. I am good at working out difficult math problems. 6. I believe that I can be successful in my math class. 8. I am confident that I can understand the material in my math class. i. I know I can learn the materials in my math class.
Achievement identity scale	Fall and spring student surveys	Students identifying and holding a self-concept as someone who can achieve academically; this student belief can improve with intervention or is a strong predictor of future math achievement	When reading the following statements, think about your current math class and decide how well the statements describe you. How much do you disagree or agree with the statements below? a. I usually do well in math. b. Math is harder for me than any other subject. c. My teacher tells me I am good at math. How much do you disagree or agree with the statements below? a. My classmates think I am good at math. b. My friends think I am good at math. c. My parents think I am good at math.
Growth mindset scale	Fall and spring student surveys	Students' belief that their ability to learn is not fixed but can be developed over time; this is a mindset that can be nurtured in instructional settings	When reading the following statements, think about your current math class and decide how well the statements describe you. How much do you disagree or agree with the statements below? a. Being a top math student requires a special talent that just can't be taught. b. If you want to succeed in math, hard work alone just won't cut it; you need to have a natural gift or talent. c. When you have to try really hard in math in school, it means you can't be good at math. d. Being a "math person" or not is something that you really can't change. Some people are good at math and other people aren't.

* The letters in "Survey items included" represent the actual survey item letters from the student and teacher surveys.

Appendix F:
Final revised AIM codebook

This page has been left blank for double-sided copying.

Core AIM Instructional Practice: Real-world mathematical inquiry and problem solving (RWMI)		
T_RWMI1	Facilitate real world inquiry	TEACHER poses a mathematical question, problem, or task with explicit real-world implications or that requires applying real-world data or information to solve.
S_RWMI1	Initiate real world inquiry	STUDENT(S) poses a mathematical question/problem or task with explicit real-world implications or that requires applying real-world data or information to solve.
S_RWMI2D	Discuss real world problem or data	STUDENT(S) discuss a mathematical question/problem, data or information with explicit real-world implications or that requires applying real-world data or information to solve
S_RWMI2P	Participate in real world inquiry	STUDENT(S) participate in a math task with explicit real-world implications or that requires applying real-world data or information to solve.

Core AIM instructional practice: Multiple representations of mathematics (MRM)		
T_MRM1	Model multiple representations	TEACHER thinks out loud to demonstrate the kinds of questions students should ask themselves to reason or make sense of different symbolic, textual, or graphical representations of mathematical concepts or relationships OR to share their rationale or justification for different solution paths.
T_MRM2	Explore multiple representations	TEACHER probes, asks purposeful questions, or provides instructions for a math task that encourages students to share, discuss or demonstrate (1) their reasoning and sense making about different symbolic, textual, or graphical representations of mathematical concepts or relationships, (2) connections or relationships of the mathematical concepts, procedures, or tasks at hand with other mathematical ideas (e.g., presented in a different lesson), or (3) alternative solution paths.
S_MRM2	Explore multiple representations	STUDENTS share, discuss or demonstrate (1) their reasoning and sense making about different symbolic, textual, or graphical representations of mathematical concepts or relationships, (2) connections or relationships of the mathematical concepts, procedures, or tasks at hand with other mathematical ideas (e.g., presented in a different lesson), or (3) alternative solution paths with other students.

Core AIM instructional practice: Mathematical discourse (MD)		
T_MD1	Model the use of math terminology	TEACHER explicitly models, reviews, or prompts students to use math terminology, typically terms that are specific to the observed lesson or instructional unit.
S_MD1	Use of math terminology	STUDENT(S) use math terminology, typically terms that are relevant to the observed lesson or instructional unit.
T_MD2	Use of common, non-technical language	TEACHER uses non-math-specific vocabulary or verbal shorthand to discuss mathematical concepts or procedures.
S_MD2	Use of common, non-technical language	STUDENT(S) uses non-math-specific vocabulary to discuss mathematical concepts or procedures.
T_MD2	Developing a collective understanding	TEACHER probes, asks purposeful questions, or provides instructions to engage more than one student to (1) evaluate or compare each other’s representations, solutions, approaches, or arguments, (2) debate math ideas and strategies, or (3) co-construct strategies or explanations in response to a mathematical task.

Core AIM instructional practice: Mathematical discourse (MD)		
S_MD2	Developing a collective understanding	More than one STUDENT (in large, small, or peer pair groups) (1) evaluate or compare each other's representations, solutions, approaches, or arguments, (2) debate math ideas and strategies, or (3) co-construct strategies and explanations in response to a mathematical task.

Core AIM instructional practice: Multilingual learner support and scaffolding (ELSS)		
T_ELSS1	Use of English language scaffolding strategies	TEACHER uses an English-language scaffolding strategy or provides linguistic support to make a math-related conversation or task more accessible. This code does not presume English as the predominate language; it refers to supporting students' linguistic understanding and fluency regardless of language.
S_ELSS2	Requests translation support	MULTILINGUAL STUDENT asks a TEACHER for language support such as what an English word or math term means or how to say something in English.
S_ELSS3	Peer language support	STUDENT(S) asks or offers translation support to another student or students engage in on-task conversation in a language other than English.

Core AIM instructional practice: Engaged student and community funds of knowledge (FoK)		
T_FoK1	Cultural funds of knowledge	TEACHER connects or employs students' community, cultural or linguistic knowledge that is specific to their individual lived experience or local context with a math-related discussion or task.
S_FoK1	Cultural funds of knowledge	STUDENT(S) connects or employs community, cultural or linguistic knowledge that is specific to their individual lived experience or local context with a math-related discussion or task.

Core AIM instructional practice: Interdisciplinary connections (IC)		
T_IC1	Make interdisciplinary connection	TEACHER explicitly connects a math-related discussion or task to another academic discipline or content area (e.g., science, social studies, art) as a tool to broaden students' understanding and application of a mathematical fact, concept, or procedure beyond the lesson.
S_IC1	Make interdisciplinary connection	STUDENT(S) connects a math-related discussion or task to another academic discipline or content area (e.g., science, social studies, art) as a tool to broaden students' understanding and application of a mathematical fact, concept, or procedure beyond the lesson.

Core AIM instructional practice: Empowered mathematical inquiry and decision making (EMI)		
T_EMI1	Facilitate empowered mathematical inquiry	TEACHER poses a question, initiates a discussion, or assigns an instructional task that requires students to use math to investigate or critique a societal challenge or a social justice issue of direct relevance to them or of their own choosing.
S_EMI1	Engage in empowered mathematical inquiry	STUDENT(S) use math to investigate or critique a societal challenge or a social justice issue of direct relevance to them or of their own choosing.

Relational interactions (RI)		
T_RI1P	Addressing student behavior	TEACHER praises student(s)' *positive* non-math-related or on-task behavior.

Relational interactions (RI)		
T_RI1N	Addressing student behavior	TEACHER redirects or reprimands student(s)' *negative, noncompliant, or off-task* non-math student behavior.
T_RI2P	Framing mathematics ability	TEACHER makes a comment that *positively* frames one or more students' general capabilities in mathematics or ability to complete an upcoming math task. Instances must include broad statements rather than a specific assessment of a contribution during the lesson.
T_RI2N	Framing mathematics ability	TEACHER makes a comment that *negatively* frames one or more students' general capabilities in mathematics or ability to complete an upcoming math task. Instances must include broad statements rather than a specific assessment of a contribution during the lesson.
S_RI2P	Framing mathematics ability	STUDENT makes a comment that *positively* frames their own or another student's general capabilities in mathematics or ability to complete an upcoming math task.
S_RI2N	Framing mathematics ability	STUDENT makes a comment that *negatively* frames their own or another student's general capabilities in mathematics or ability to complete an upcoming math task.
T_RI4P	Setting the emotional tone	TEACHER sets *positive* expectations for the classroom culture/climate by preempting behavioral issues with compassion and empathy or creating a safe emotional space for students.
T_RI4N	Setting the emotional tone	TEACHER sets *negative* expectations for the classroom culture/climate by preempting behavioral issues with threats, warnings or other statements of *negative* consequences.
T_RI5	Scaffolding discourse	TEACHER provides math-related feedback, asks questions, or models the thinking process to help a student break down a cognitively demanding or complex task into more manageable, accessible, or comprehensible parts.
S_RI6T	Requesting assistance	STUDENT asks a teacher for math-related help with a lesson-related activity that advances their understanding of a math concept or ability to complete a mathematical procedure.
S_RI6S	Requesting assistance	STUDENT asks *another student* for math-related help with a lesson-related activity that advances their understanding of a math concept or ability to complete a mathematical procedure.
T_RI7P	Valuing math persistence and a growth mindset	TEACHER encourages students to work through cognitively demanding tasks by *praising* confusion and mistakes or encouraging productive struggle.
T_RI7N	Devaluing math persistence	TEACHER discourages working through cognitively demanding tasks by *reprimanding or ridiculing* struggle, confusion, and mistakes.
T_RI7D	Discomfort with productive struggle	TEACHER demonstrates discomfort with one or more students struggling to complete an instructional task by jumping in to help shortly after assigning a task.
S_RI7P	Valuing math persistence	STUDENT expresses about themselves or encourages others to work through cognitively demanding tasks by *praising* struggle, confusion, and mistakes.
S_RI7N	Devaluing math persistence	STUDENT expresses about themselves or discourages others from working through cognitively demanding tasks by *reprimanding or ridiculing* struggle, confusion, and mistakes.

Relational interactions (RI)		
T_RI8M	Correcting	Correcting a student's *math-related* misconceptions, error, or misstep by sharing the correct answer or demonstrating the appropriate approach.
T_RI8NM	Correcting	Correcting *non-math-related* errors (e.g., grammar, pronunciation, vocabulary)
T_RI9	Moderating the amount of speech	TEACHER urges student(s) to speak less or more when discussing math-related ideas or content.
T_RI10	Rhetorical questioning	TEACHER asks a rhetorical math-related question for which they do not expect a response.
T_RI11	Non-inclusive instructional decision	TEACHER makes an instructional decision that could be perceived as unrelatable, problematic, or inappropriate by one or more students in the classroom.
T_RI12P	Giving Affirming Feedback	Teacher gives a student positive, supportive, or constructive feedback on their math-related work or contributions—but does not elaborate or explore as to why the work is good.
T_RI12N	Giving Negative Feedback	Teacher gives a student negative, unconstructive, or unsupportive feedback on their math-related work or contributions—but does not explain why the work is poor.
T_RI12NT	Giving Neutral Feedback	Teacher gives a student feedback that does not evaluate, confirm or refute the accuracy of their answer. The teacher simply acknowledges that the student has offered a response or made a contribution.
T_RI13	Interpersonal connection	TEACHER forges or reinforces a personal or relational connection with one or more students via a shared interest, expressing curiosity or appreciation for a student's interest, or engaging with a student in their home language.

Procedural practice (PP)		
T_PP1	Taking attendance	TEACHER verbally or nonverbally takes attendance, counts students, or otherwise indicates that they are taking note of present/absent students.
T_PP2	Collecting homework/classwork	TEACHER physically or digitally collects student work.
T_PP3	Assigning homework/classwork	TEACHER assigns a homework or classwork assignment.
T_PP4	Making an announcement	TEACHER makes an announcement that is not related to the current math lesson.
T_PP5	Establishing or reinforcing classroom norms	TEACHER explains, discusses, or reminds students of classroom procedures, rules, or code of conduct that is not specific to the math lesson.
T_PP6	Initiation-Response-Evaluation (IRE) questioning	TEACHER poses a question—for which there is a presumption of a "correct" or specific answer and that requires no elaboration or justification on the student's part—assesses the correctness of a student's response, and gives close-ended feedback such as a yes/no.
T_PP7	Lecturing or demonstrating	TEACHER presents, demonstrates, reviews, defines, summarizes, or introduces instructional content in a non-interactive manner for an extended period of time.
T_PP8	Procedural clarification	TEACHER provides a clarification or reminder about the instructions for a lesson-related activity that has already been assigned.

Procedural practice (PP)		
S_PP8	Procedural clarification	STUDENT(S) asks the teacher a procedural, non-math related question to clarify expectations for an activity.
S_PP9	Warm up/close out	STUDENT(S) complete a brief and procedurally normed activity at the opening or closing of the class.

Performance tasks (PT)		
S_PT1	Memorize or recall	STUDENT(S) commit to memory or reproduce previously learned facts, rules, formulas, or definitions without connection to the concepts or meaning that underlie.
S_PT2	Perform procedures	STUDENT(S) use an algorithm or procedure to solve a problem with a focus on producing correct answers. No explanation is required, or explanations focus solely on describing the procedure that was used.
S_PT3	Demonstrate understanding	STUDENT(S) focus on the use of procedures for the purpose of developing understanding of mathematical concepts and ideas or providing explanations for why steps in a procedure make sense.
S_PT4	Conjecture, generalize, or prove	STUDENT(S) notice patterns or make observations and use these to form a conclusion; they engage in complex, non-algorithmic thinking to explore and understand the nature of mathematical concepts, processes, or relationships.
S_PT5	Solve non-routine problems or making connections	STUDENT(S) use relevant knowledge and experiences to work through a novel task or a task that could be represented or solved in multiple ways; student makes connections among various representations or strategies.

Grouping (G)		
G1	Whole class	TEACHER facilitates an instructional task, discussion, or presentation to the entire class.
G2	Small group	TEACHER assigns an instructional task to one or more students to be completed in small groups (groups of 3-8 students) based on proximity, classroom norms, or student choice.
G3	Pair	TEACHER assigns an instructional task to one or more students to be completed in pairs (groups of 2 students).
G4	Individual	TEACHER assigns an instructional task to one or more students to be completed independently (in isolation from or without support or collaborating with other students).
G5	Ability or strategic grouping	TEACHER assigns students to groups to personalize or differentiate the math learning environment such as by learning need, learning preference, or ability.
G6	Stations	TEACHER assigns students to different learning stations, typically distinguishable by different learning activities and locations within the classroom.

Instructional materials (IM)		
IM1	Textbook/Workbook	TEACHER or STUDENTS interact with a textbook or associated workbook manufactured by a curriculum company.

Instructional materials (IM)		
IM2P	Worksheet/handout: Paper-based	STUDENTs interact with a structured paper-based document with instructions, tasks, and space for students to complete work.
IM2E	Worksheet/handout: Electronic	STUDENTS interact with a structured electronic document with instructions, tasks, and space for students to complete work.
IM3	Blackboard/whiteboard/smartboard/overhead	TEACHER or STUDENTS interact with a large board or screen, visible to all students in the room, to facilitate whole-class learning. This may include a document camera or transparency machine.
IM4	Audio-visual recording	TEACHER or STUDENTS interact with a video clip, audio clip, or digital timer.
IM5I	Computer/Tablet: Individual	STUDENTS interact with individual or personal devices to access curriculum content, complete activities, or submit work.
IM6P	Assessment: Paper-based	STUDENTS interact with paper-based assessments or tools for teachers to gauge student learning.
IM6C	Assessment: Electronic	STUDENTS interact with electronic assessments or tools for teachers to gauge student learning.
IM7	Learning management system or other educational technology	TEACHER or STUDENTS interact with *or reference* a digital learning management system or another educational technology tool.
IM8A	Manipulative: Analog	STUDENTS interact with physical objects which support learning or engagement with a specific math concept.
IM8D	Manipulative: Digital	STUDENTS interact with digital tools which support learning or engagement with a specific math concept.
IM9	Unstructured materials	STUDENTS interact with blank or unstructured materials with no scaffolding or written structure.
IM10	Other	TEACHER or STUDENTS interact with or reference any other instructional material. Note what it is in the running records.

Instructional material type (IMT)		
IMT1	Core curriculum	The primary textbook the teacher is instructed to use by the school or district. The core curriculum should be one of the 6 study curriculums.
IMT2SD	Supplemental: Curriculum/learning platform developer	Content or materials developed by a curriculum or learning platform developer that is not part of the core curriculum. This may include purchased or free materials.
IMT2SS	Supplemental: State or district developed	Content or materials developed by the teacher's state or district that is not part of the core curriculum. This may include pacing charts or guidance about standards to prioritize.
IMT2ST	Supplemental: Teacher developed	Content or materials developed by the teacher observed or by another teacher.
IMT2SO	Supplemental: Other	Any other content or materials developed by a source not captured in an above code.
IMT3C	Culturally Responsive	Content or material that incorporates culturally responsive content or is a culturally responsive artifact—whether or not it furthers math learning.
IMT3L	Language aid for multilingual learner	Content or material that has been adapted to support multilingual learners.

Appendix G.
Central tendency and variability of the AIM teacher performance scales

This page has been left blank for double-sided copying.

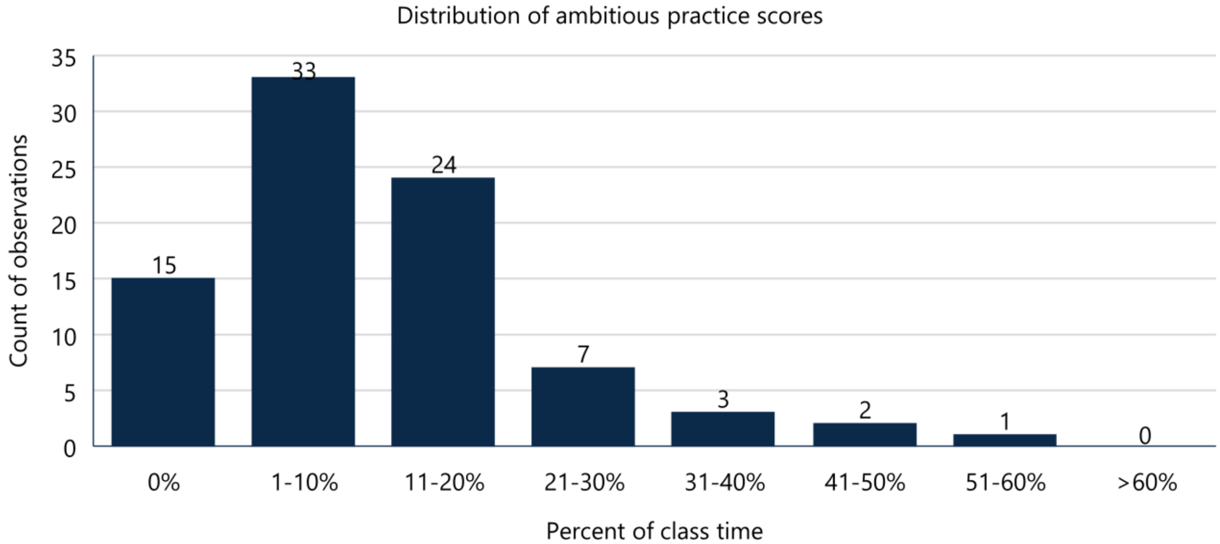
After estimating the reliability and validity for the AIM tool, we explored the extent to which the tool can be used to distinguish practice across instructional contexts (such as district, grade, teacher characteristics, and student demographics) and curricula. We ran descriptive statistics to explore the central tendency and variability of the AIM performance scales (Exhibit G.1).

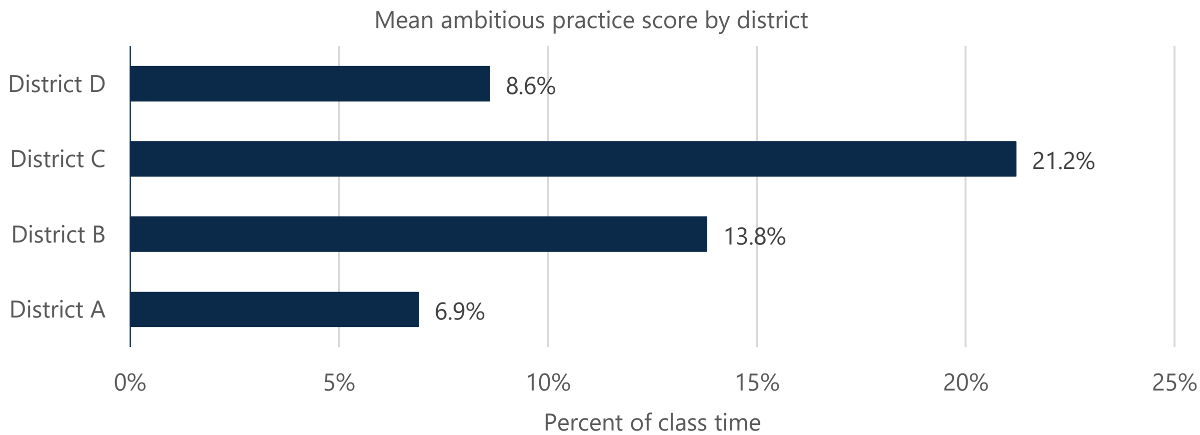
Exhibit G.1. Descriptive statistics for the AIM teacher performance scales

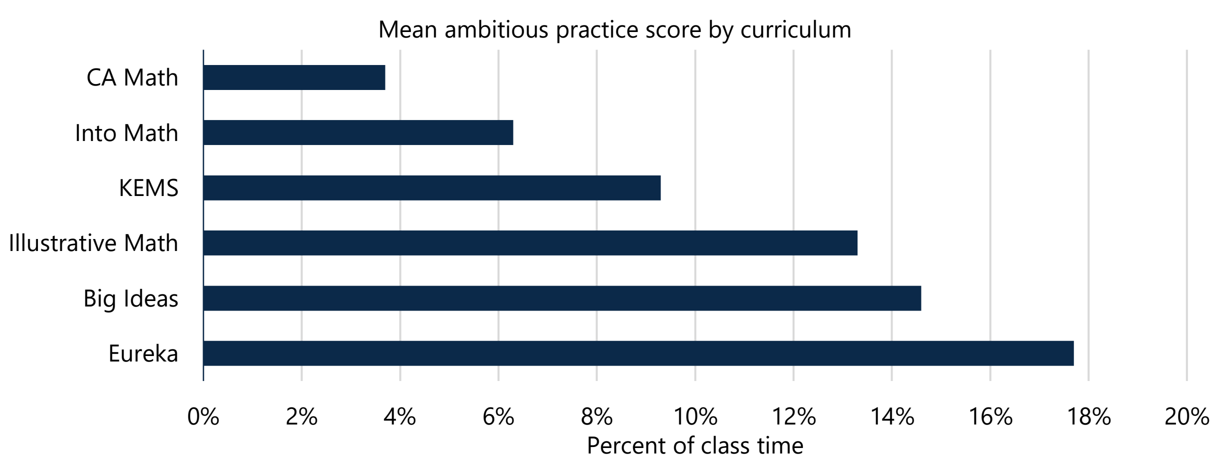
	n	Min	Max	Mean	Std. deviation
Ambitious practice	85	0.0%	62.5%	13.2%	11.4%
Inclusive practice	85	1.5%	42.7%	11.3%	8.7%
Core AIM instructional practice	85	0.4%	45.6%	9.5%	8.8%
Student-centered practice	85	0.6%	46.1%	11.2%	9.3%
Teacher-centered practice	85	2.8%	78.1%	23.4%	13.5%

Below, we share visualizations of the variability of AIM teacher performance scale scores.

Exhibit G.2. Variability of the AIM ambitious practice scale



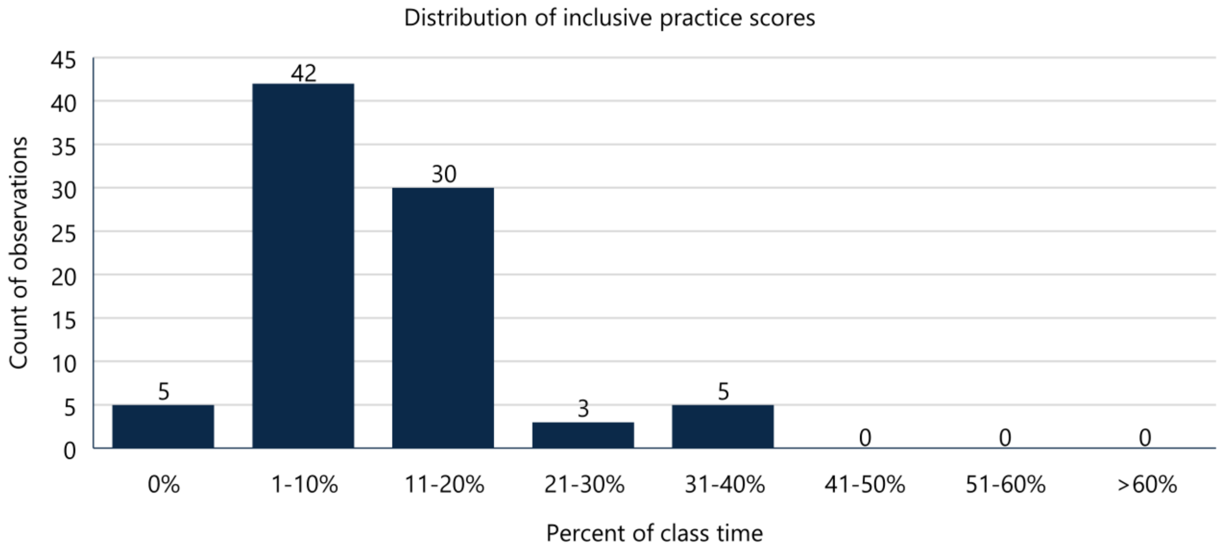


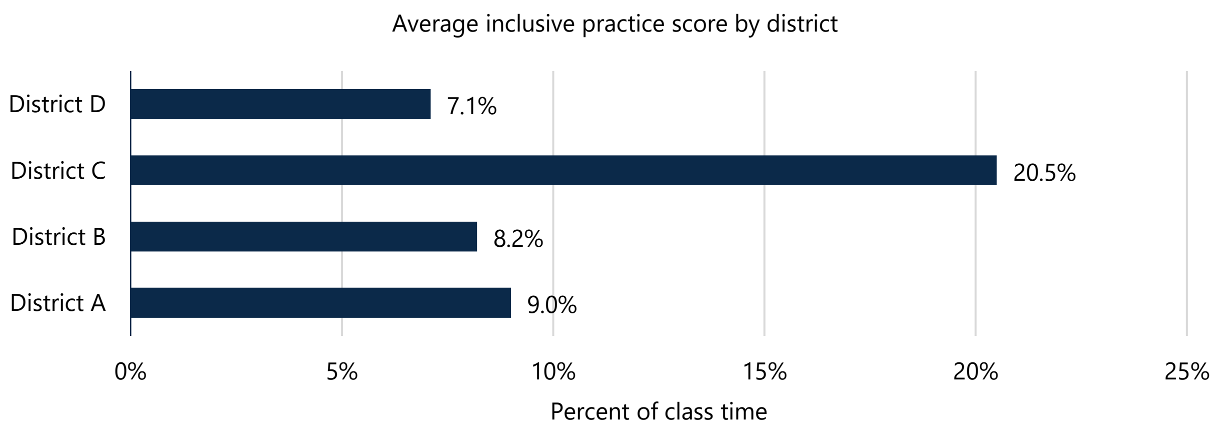


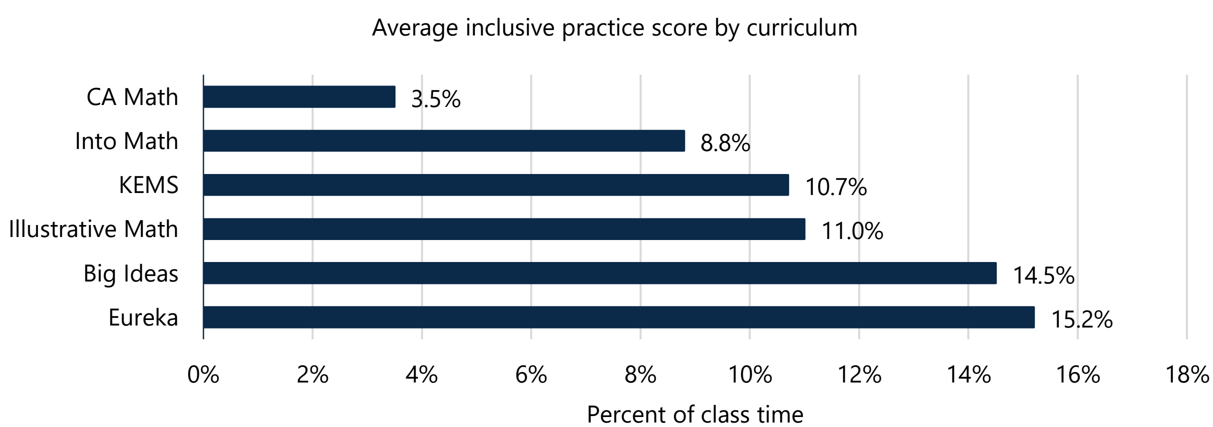
Source: SY 2021–2022 and SY 2022–2023 AIM classroom observation data.

n = 85 observations.

Exhibit G.3. Variability of the AIM inclusive practice scale

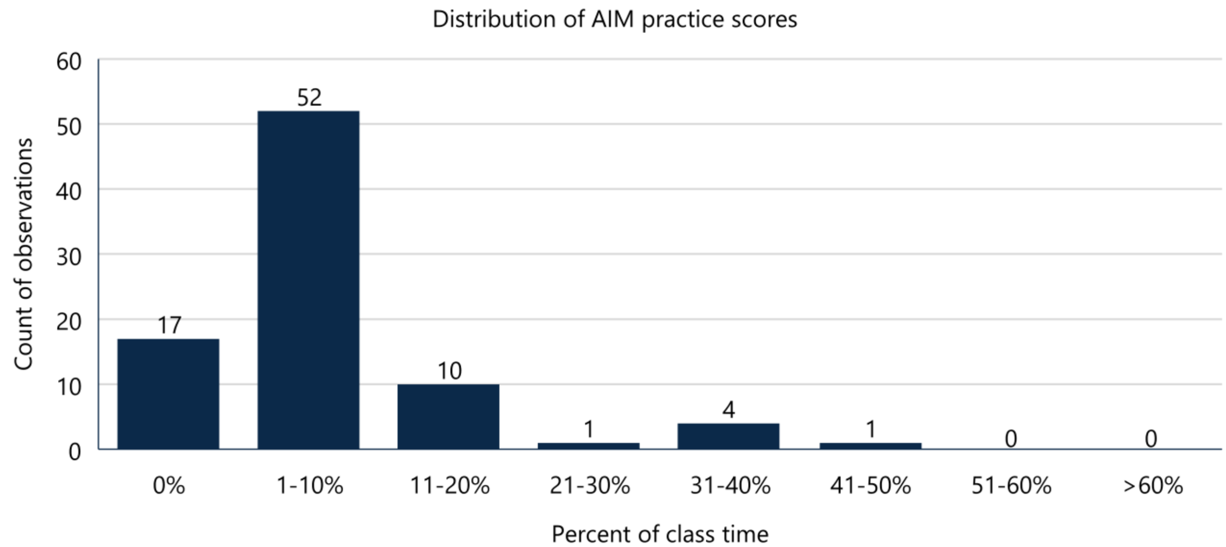


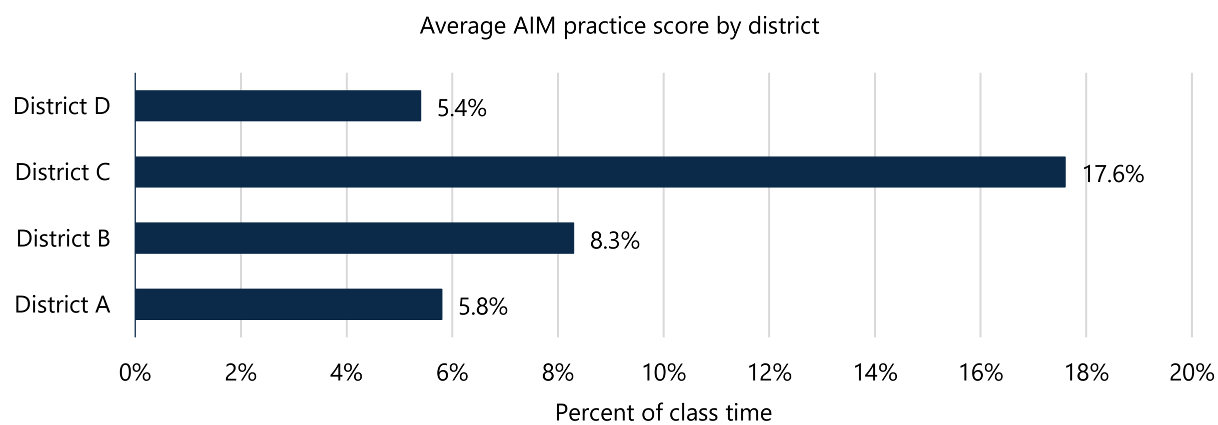


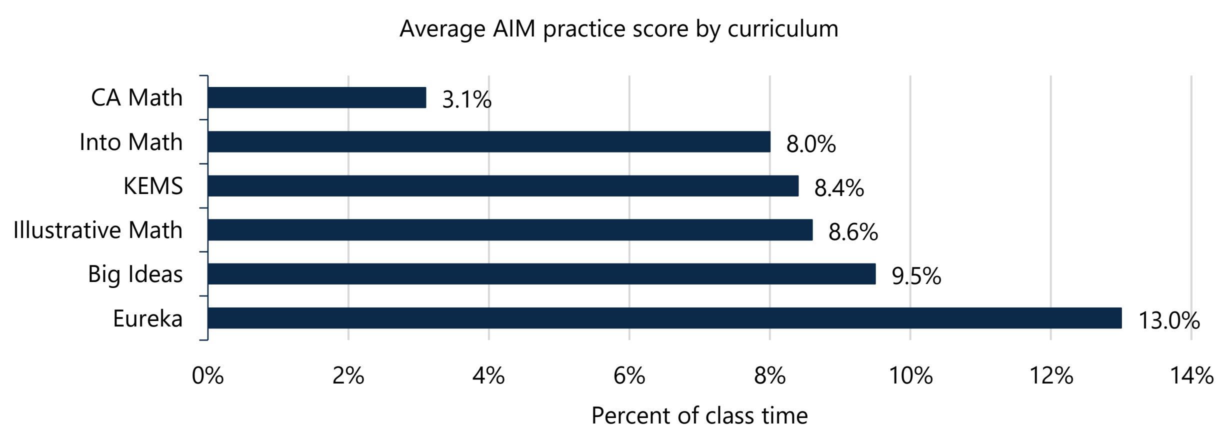


Source: SY 2021–2022 and SY 2022–2023 AIM classroom observation data.
n = 85 observations.

Exhibit G.4. Variability of the core AIM instructional practice domains



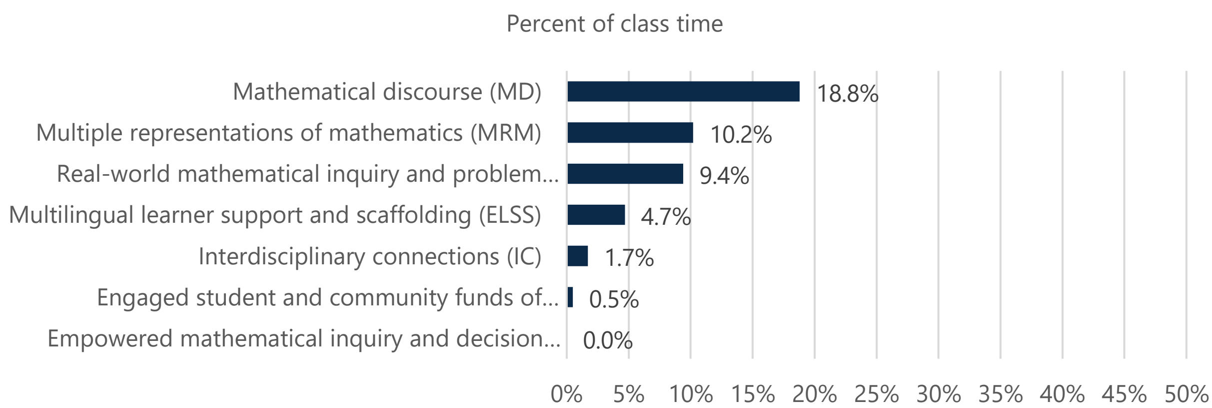




Source: SY 2021–2022 and SY 2022–2023 AIM classroom observation data.
n = 85 observations.

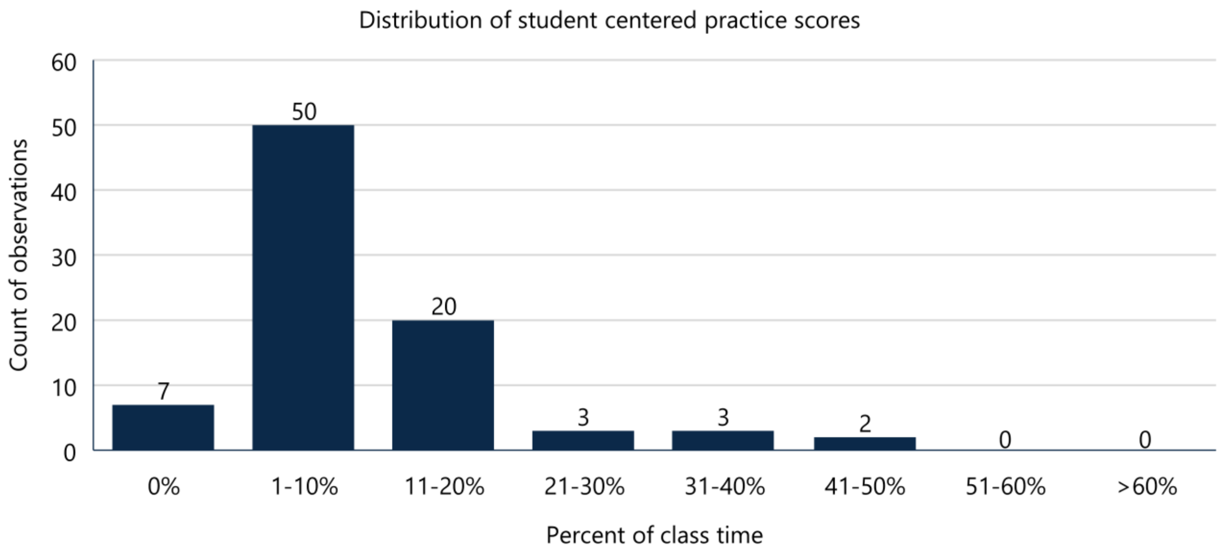
It should be noted that, of the teaching strategies that comprise the scale scores, strategies most commonly associated with culturally responsive teaching in mathematics were rarely used. On average, just 0.5 percent of the instructional time we observed engaged students' cultural and community funds of knowledge. We did not observe any instances of teachers creating opportunities for students to use mathematics to investigate social justice issues (Exhibit G.5)

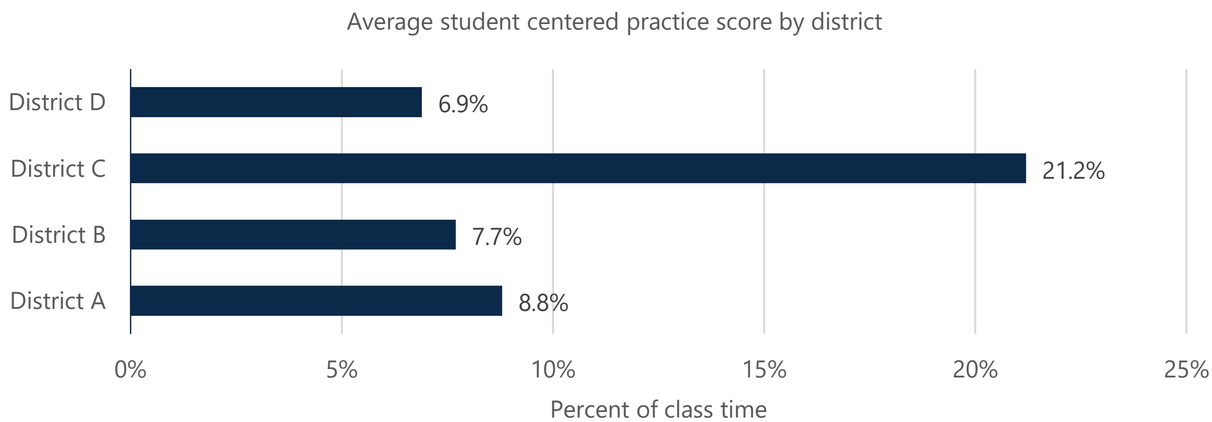
Exhibit G.5 Use of AIM instructional strategies

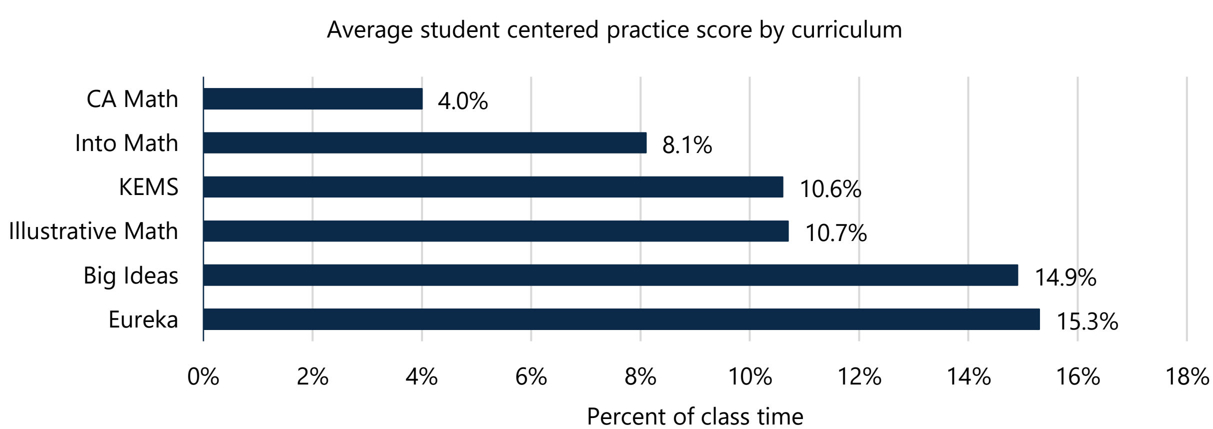


Source: SY 2021–2022 and SY 2022–2023 AIM classroom observation data.
n = 85 observations.

Exhibit G.6. Variability of the AIM student-centered practice scale

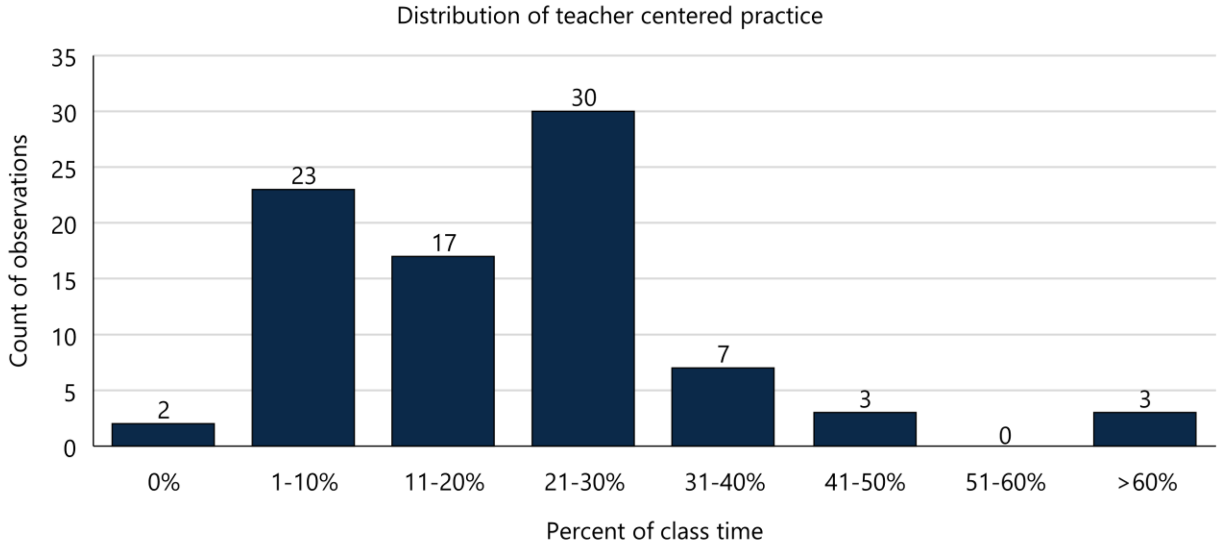


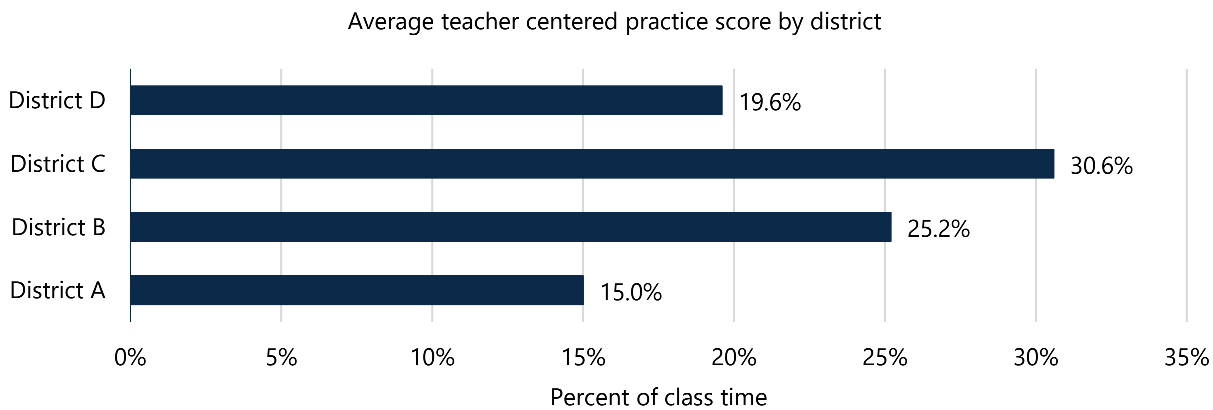


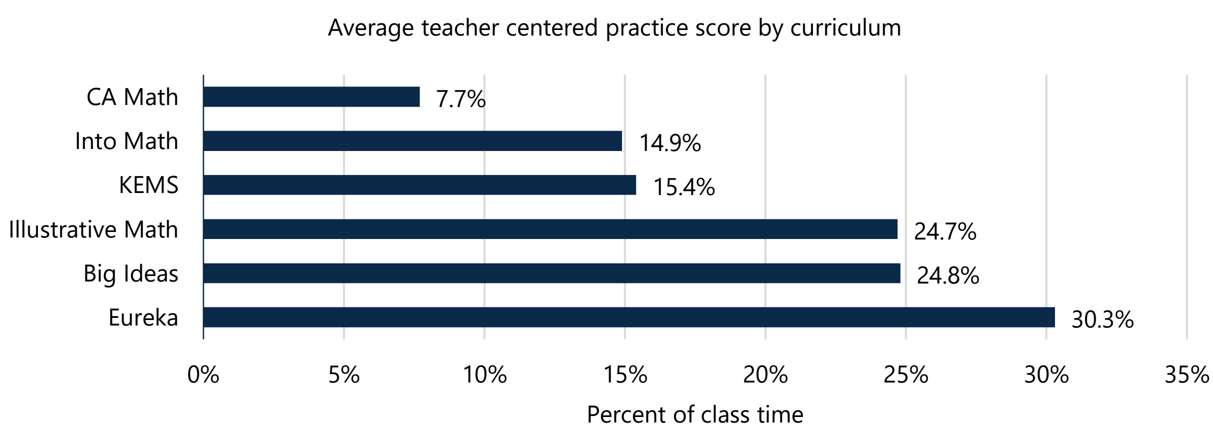


Source: SY 2021–2022 and SY 2022–2023 AIM classroom observation data.
n = 85 observations.

Exhibit G.7. Variability of the AIM teacher-centered practice scale







Source: SY 2021–2022 and SY 2022–2023 AIM classroom observation data.
n = 85 observations.