Julie Bruch, Ijun Lai, and Veronica Sotelo Muñoz

Menu of Measures: High-Quality Measures of Middle Years Math Student Outcomes

Measuring the impacts of math instruction and interventions is an important part of learning how to improve math performance. Identifying and administering high-quality measures will help the field understand how to improve students' knowledge and enjoyment of math.

This is a menu of measures, designed to be a resource to help educators, districts, researchers, and funders identify measures to learn about student math knowledge and enjoyment in a valid, reliable, and accessible way. The menu is not a comprehensive list of every math measure available in the field. It is a curated list of measures developed in partnership with a panel of researchers in the field of math instruction and measurement. The menu includes measures that meet some or all of a set of preferred criteria developed by the panel (see Appendix A at the end of the menu for full criteria). It includes a list of measures in two student outcome areas:

1. **Math knowledge,** measuring the extent to which students develop a deep knowledge of math and can flexibly and accurately solve mathematical problems, and provide justifications for their approach
2. **Math enjoyment,** measuring students' attitudes and mindsets related to math (see Table 1 for definitions of specific math enjoyment constructs covered in this menu)

For each measure, we provide a brief summary of the measure, key publications, reliability information, the student outcome constructs that the measure covers, and other considerations for use.

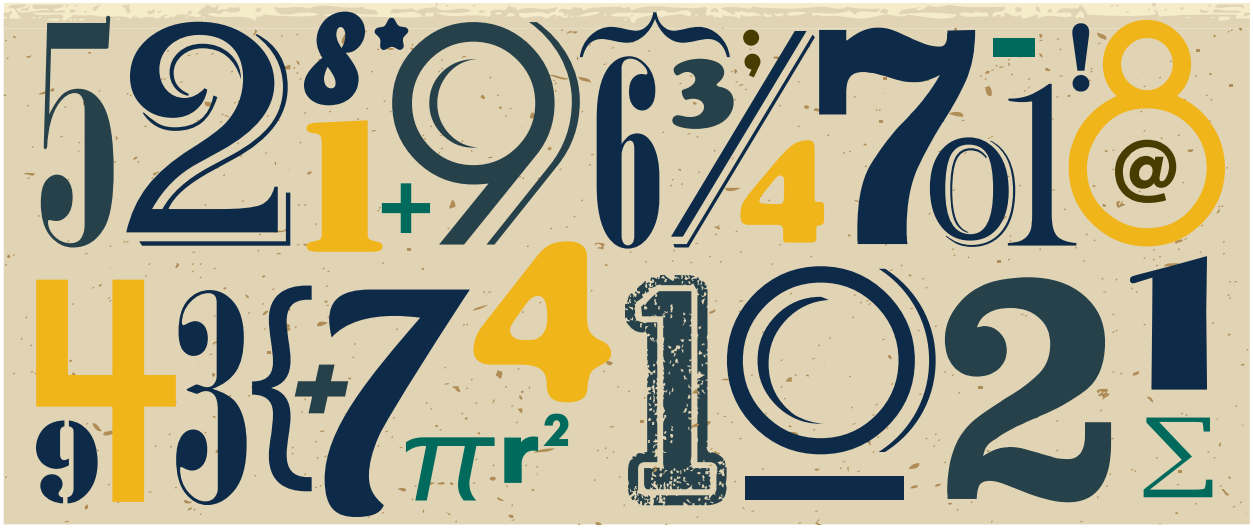


Table 1. Math enjoyment constructs

Construct	Definition
Sense of belonging	A student's feeling of being connected to the math learning content and activities and being an accepted, respected, and valued math learner
Self-efficacy/confidence	Self-efficacy is defined as a student's perception of his or her ability to successfully complete a specific math task, and can vary by task. Math confidence is a student's global assessment of his or her math ability
Enjoyment/general attitude	A student's interest, liking, valuing, or enjoyment of mathematics as well as affective reactions to and attitudes toward mathematics
Growth mindset	A student's belief that math abilities can be developed and improved through hard work, practice, good strategies, and input from others (as opposed to the belief that math abilities are fixed, stable, and unable to change)
Value and importance of math	A student's perception about the importance of doing well on mathematics and usefulness of mathematics for many aspects of daily life or fulfilling future goals

Note: An additional construct, positive math identity, overlaps with some of the other constructs in this table.

The menu includes the following measures:

Math knowledge

- K1. ClearSight
- K2. Performance Series
- K3. PSAT 8/9
- K4. Trends in International Mathematics and Science Study
- K5. Mathematics Assessment Collaborative/ Mathematics Assessment Resource Service (MAC/MARS)
- K6. Star Math
- K7. Iowa Algebra Aptitude Test
- K8. Test for Understanding Fractions

Math enjoyment

- E1. Attainment Value of Mathematics
- E2. Beliefs, Engagement, and Attitude of Math Motivation Scale
- E3. Expectancy-Cost-Value Scale
- E4. Indiana Mathematics Belief Scale
- E5. Math and Me Survey
- E6. Math and Science Engagement Scales
- E7. Math Identity Measures
- E8. Mathematical Mindset
- E9. Mathematics Attitude Inventory
- E10. Mathematics Self-Efficacy Scale
- E11. Sense of Connectedness to My Mathematics Classroom ◢

Assessment of measures

Measures are assessed based on:

- **Reliability.** What is known about whether scores are consistent across individuals, context, time, items, or raters?
- **Validity.** Do the scores represent what they intend to measure?
- **Cultural responsiveness.** Is the measure reliable and valid for use in classrooms that include students who are Black, Latino, and/or experiencing poverty? For instruments that students complete, are the items relevant for students' sociocultural values and experiences?
- **Linguistic accessibility.** For instruments that students complete, are there appropriate adaptations, translations, and resources available for students who need assistance with written or oral English language?
- **Feasibility to administer.** Are instruments feasible to administer and score, and are there few barriers in terms of cost and licensing? Can the instruments be administered virtually, as might be required in the 2021-2022 school year?

All measures in the menu are reliable and specific to math. Reliability must be greater than 0.70 for standardized tests that have multiple items related to the same construct; surveys with self-reported measures of belief, attitudes, and behaviors related to the same construct; and measures that have more than one rater assigning ratings or scores. For some math knowledge assessments, developers assert that assessments are reliable, but data are not publicly available. For some survey measures, the overall instrument has shown to be reliable, but there could be individual subscales that are not. In these cases, users should not administer subscales in isolation.

When selecting measures, users can consider the extent to which measures meet the preferred criteria. Preferred criteria (detailed in Steps 2 and 3 of Appendix A) include reliability and validity with students who are Black, Latino, and/or experiencing poverty; availability of the instruments themselves and guidance to consistently interpret results; the linguistic accessibility of the measures; and the extent to which the measures are appropriate to use in a virtual setting. There are also content-specific criteria for mathematics knowledge (the extent to which the measures deep knowledge, specific math proficiency strands, and cognitively complex concepts) and enjoyment (alignment with key enjoyment constructs and the extent to which the measure state-based instead of trait-based characteristics).

Tables 2 and 3 show the specific preferred criteria that each measure meets. Each column in the table lists an assessed measure, and each row describes a relevant criterion that was examined; the intersecting cell presents the panel's assessment of whether each measure meets the criterion of interest.

When using this resource, readers should also consider a set of context-specific questions to determine whether the recommended measures are a good fit for the local context (see next page). The list of measures includes a brief description of each measure, key publications that describe or test the measure, and other considerations for implementation and interpretation.

Consider context-specific questions to determine fit of measures

Questions to assess validity:

- Is the measure designed to capture an outcome in the program's theory of change?
- Is the measure aligned with an outcome expected to change?
- Is there evidence that the measure is predictive of expected longer-term outcomes?
- Is the measure designed to be a formative or summative assessment?
- Does the measure capture growth or proficiency?

Questions to assess context-specific linguistic accessibility and cultural relevance:

- Is the wording and content relevant and appropriate for the students' level of literacy and cognition? If not, can the measure be adapted to be developmentally appropriate?
- Are there languages other than English that the participating students speak? If so, is the measure linguistically accessible in all relevant languages and dialects?
- Is the interpretation of measure items relevant to students' sociocultural values and experiences and does not presume White middle-class values?

Consider context-specific questions to determine usability of measures

- Is it feasible to administer the measure given training requirements?
- Is it feasible to administer the measure given scoring requirements?
- Is it feasible to administer the measure given the cost?
- Is it feasible to administer the measure given the time required to administer? ◢

Table 2. Math knowledge

Criterion	K1. ClearSight	K2. Performance Series	K3. PSAT 8/9	K4. Trends in International Mathematics and Science	K5. MAC/MARS	K6. Star Math	K7. Iowa Algebra Aptitude Test	K8. Test for Undersating Fractions
Grades tested in prior research[1]	K-12	K-12	8-9	4, 8, 12	3-10	1-12	7-8	4-5
Item availability	IB	IB	W	Sub	W	W, IB*	W	W
Demonstrated reliability and validity with priority communities	No	No	No	Yes	No	Yes	No	No
Measure is available for use without restrictions on access	No	No	No	Yes	No	No	No	Yes
Guidance available to help users consistently interpret results	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Measure is linguistically accessible	No	Yes	No	Yes	No	Yes	Yes	Yes
Measures can be implemented in virtual setting	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Assesses deep knowledge of math	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Assesses one of the math proficiency strands[2]	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Assesses cognitively complex concepts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

No = The panelists determined that there is not sufficient evidence to determine whether the measure meets the criterion. This could mean either that the studies reviewed did not report the necessary evidence OR that the instrument is not available for assessment by the panel.

Item availability indicates the ability for programs to use individual items or subscales from a larger assessment. IB = item bank; programs may be able to select individual items that are most applicable. Sub = subscale; programs may be able to use scores from individual subscales that are most applicable; W = whole test; programs must use a comprehensive score and cannot customize the assessment.

*Star Math is marked as both whole test and item bank because there are two versions of the assessment. One version is narrower, and does not allow for customization, while the custom version allows programs to customize the tests.

[1] Prior research refers to the key publication reviewed by panelists to assess each measure. The grade range indicates the grades included in each study. For standardized tests, we list grades with available tests.

[2] Math proficiency strands include procedural fluency (ability to carry out procedures flexibly, accurately, efficiently, and appropriately), conceptual understanding (fundamental understanding of mathematical ideas and the ability to integrate new ideas into their understanding by connecting with previous knowledge), strategic competence (ability to formulate, represent, and solve mathematical problems), and adaptive reasoning (capacity for logical thought, reflection, explanation, and justification).

Table 3. Math enjoyment

Criterion	E1. Attainment Value of Mathematics	E2. Beliefs, Engagement, and Attitude of Math Motivation Scale	E3. Expectancy-Cost-Value Scale	E4. Indiana Mathematics Belief Scale	E5. Math and Me Survey	E6. Math and Science Engagement Scales	E7. Math Identity Measures	E8. Mathematical Mindset	E9. Mathematics Attitude Inventory	E10. Mathematics Self-Efficacy Scale	E11. Sense of Connectedness to My Mathematics Classroom
Included in list of highly recommended measures	No	No	Yes	No	Yes	Yes	No	No	No	No	No
Grades tested in prior research[1]	5-9	2-3	1-PS	PS	3-6	5-12	PS	6-8	7-12	9- 10	5-9
Demonstrated reliability and validity with priority community	Yes	No	No	No	No	Yes	No	No	No	No	Yes
Measure is available for use without restrictions on access	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Guidance available to help users consistently interpret results	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Measure is linguistically accessible	No	No	No	No	No	Yes	No	No	No	No	Yes
Measures can be implemented in virtual setting	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Assesses at least one key math enjoyment construct[2]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Assesses state-based instead of trait-based characteristic[3]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

No = The panelists determined that there is not sufficient evidence to determine whether the measure meets the criterion. This could mean either that the studies reviewed did not report the necessary evidence OR that the instrument is not available for assessment by the panel.

Underlined instruments are highly recommended.

PS = postsecondary.

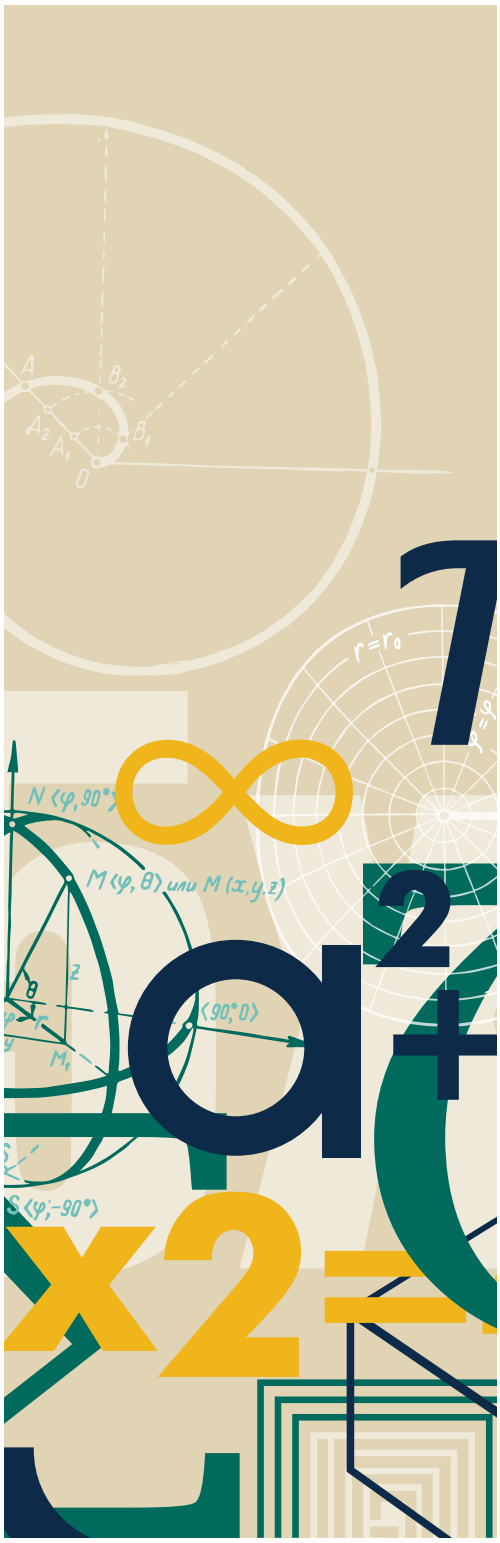
[1] Prior research refers to the key publication reviewed by panelists to assess each measure. The grade range indicates the grades included in each study.

[2] The panel identified a set of key math enjoyment constructs that are distinct from one another and can be measured. The list of constructs is in Table 1.

[3] State-based characteristic refers to a temporary way of thinking, feeling, behaving, and so on, while a trait-based characteristic is considered to be a stable and enduring characteristic.

Priority measures for the Middle Years Math grantees

The panel identified a narrow set of measures for a group of Middle Years Math grantees funded by the Bill & Melinda Gates Foundation. These grantees tested innovative solutions to improving math outcomes for students in middle grades who are Black, Latino, and/or experiencing poverty. The solutions range from tutoring initiatives to educational technologies to summer programs. For the initial set of grantees, the panel did not recommend a common assessment to measure math knowledge, and some used locally required standardized assessments. However, the panel highly recommended a common set of instruments to measure math enjoyment. These instruments measure several constructs of math enjoyment that the panel prioritized for this set of grantees, the types of interventions they implemented, and the research questions they sought to answer. The three instruments include:



- The ***Expectancy-Cost-Value Scale*** (Kosovich et al. 2014; Lauermann et al., 2017; Simpkins et al., 2006; Wang, 2012) is a 10-item student survey instrument based on the expectancy-value theory of motivation that uses a 6-point Likert scale. The survey has three subscales: expectancy (measures self-efficacy/confidence; example: "I know I can learn material in my math class."), cost (example: "My math classwork requires too much time."), and value (example: "I value my math class.").
- The ***Math and Me Survey*** (Adelson & McCoach, 2011) is a 27-item scale students in grade 3 to 6 complete to measure their beliefs on three constructs using a 5-point Likert scale. Subscales include mathematical self-perceptions (example: "Math comes easily to me."), enjoyment of mathematics (example: "Math is fun."), and perceived usefulness of mathematics (example: "Math is all around us in our everyday lives.").
- The ***Math and Science Engagement Scales*** (Wang et al., 2016; Fredricks et al., 2016) comprises 33 items (17 positively worded and 16 negatively worded) that are on a 5-point Likert scale. Subscales include behavioral engagement (example: "I put effort into learning science/math."), emotional engagement (example: "I look forward to science/math class."), cognitive engagement (example: "I think about different ways to solve a problem."), and social engagement (example: "I try to understand other people's ideas in science/math class."). ◢

Description of measures

This section describes each measure included in the menu, including additional considerations for implementation and interpretation. The information in this section can help users select a measure from among several options they are considering. It can also inform plans for implementing and interpreting specific measures.

Measures of math knowledge

Measures of math knowledge in this section assess the extent to which students have deep math knowledge instead of rote memorization of facts; develop skills in a math proficiency strand such as procedural fluency and adaptive reasoning; and understand cognitively complex concepts (see Appendix A for a detailed description of math proficiency strands).

While there are many ways to measure math knowledge (e.g. performance assessments, open responses, student portfolios), this menu focused on standardized assessments that can be used in consistent ways across different contexts and that are publicly available or easily accessible.

The majority of the measures (K1-K7) are standardized tests with proprietary test items. The panel assumes that standardized tests have face validity and are reliable. Technical manuals can be requested from developers if additional information is needed.

K1. **ClearSight (formerly known as AIRways)** is a web-based math assessment with questions aligned to grade-level state standards and is customizable for grades kindergarten through high school. Each benchmark assessment has 6 to 24 items, though no information is available about the amount of time allotted for each assessment.

- *Key publication:* No publicly available evidence; documentation may be available through direct request to the measure developer. Information about the assessment from the developer, Voyager Sopris Learning, is available here.
- *Considerations for reliability:* The developer asserts that measures are reliable and valid. Although this test has been used with students in the priority communities, reliability is not reported for racial or ethnic subgroups. Additional testing is needed with the priority communities.
- *Relationship to math knowledge:* Items appear to assess deep math knowledge, math proficiency strands, and cognitively complex concepts. Specific examples and verification of content can be requested from the developer.
- *Other considerations for use*: Brochure states that the developer can create assessments for students from kindergarten through high school, and as many interim assessments as the consumer would like. The assessment is only available through purchase.

K2. The **Performance Series** is a web-based computer-adaptive diagnostic assessment to predict student scores on state tests and can track student growth over time. The developer, Scantron, offers both online and paper/hard-copy assessments. Item bank includes questions for students in kindergarten through high school. The average time to administer the test is approximately 20 minutes.

- *Key publication:* No publicly available evidence; documentation may be available through direct request to measure developer. Information about the assessment is available from the developer here.
- *Considerations for reliability:* Reported reliability from the developer is 0.91. Although this test has been used with students in the priority communities, reliability is not reported for specific racial or ethnic subgroups. Additional testing is needed with the priority communities.
- *Relationship to math knowledge:* The public does not have access to math item banks. Test developer staff can work with grantees to determine which types of items to use. Items appear to assess deep math knowledge, math proficiency strands, and cognitively complex concepts.
- *Other considerations for use*: Participating schools need to administer assessment and share data, or grantee needs to purchase for use ($1,200 per school per year). The mathematics assessments portion is available in both English and Spanish. The Spanish version is developed through a transadaption process, ensuring that test items have the same meaning, text complexity, and other measurement characteristics as the English version (for more information, click here).

K3. The **PSAT 8/9** is a standardized math test with an emphasis on problem solving, modeling, using tools strategically, and using algebraic structure. Most questions are multiple choice, but some require students to write in an answer instead of selecting an answer. This is a 60-minute test with 38 questions and tasks.

- *Key publication:* No publicly available evidence; documentation may be available through direct request to measure developer. Information about the assessment is available from the developer here.
- *Considerations for reliability:* The developer asserts that measures are reliable and valid. Although this test has been used with students in the priority communities, no public information is available about reliability and validity for these communities. Additional testing is needed with the priority communities.
- *Relationship to math knowledge:* This test intentionally has a wide range of mathematical tasks, which are heavily based on the National Council of Teachers of Mathematics standards. Items appear to assess deep math knowledge, math proficiency strands, and cognitively complex concepts.
- *Other considerations for use*: The cost is $17 per student.

K4. The **Trends in International Mathematics and Science Study (TIMSS)** is an international test of mathematics and science achievement for students in grades 4, 8, and 12. These items are all available for use via download; grantees can select which items and number of items they would like to use.

- *Key publication:* Reliability and validity tests are analyzed after every test release, so there is more than one key publication about this test. For example, Mullis et al. (2016) evaluate the reliability and validity of the 2015 TIMSS.
- *Considerations for reliability:* No public information is available about reliability and validity for subgroups of students defined by race, ethnicity, or family income within each country, but the TIMSS has been used in more than 60 nations and in ethnically diverse school districts such as Chicago, Jersey City, and Dade County.
- *Relationship to math knowledge:* This measure is an assessment of mathematical reasoning and problem solving. Items appear to assess deep math knowledge, math proficiency strands, and cognitively complex concepts.
- *Other considerations for use*: After each round of testing, TIMSS releases half of the questions for free public use, which can be found here and are available for download. Each item has been tested for reliability and validity.

K5. The **Mathematics Assessment Collaborative/ Mathematics Assessment Resource Service (MAC/ MARS)** is a 40-minute standardized test that includes a set of detailed math performance tasks. Tests are available for use in grades 3 to 10. A new test for each grade is created each year.

- *Key publication:* No publicly available evidence; documentation may be available through direct request to the measure developer. Information about the assessment is available from the developer here.
- *Considerations for reliability:* The developer asserts that measures are reliable and valid. No public information is available about reliability and validity for students in the priority communities. Additional testing is needed with students in the priority communities.
- *Relationship to math knowledge:* This test assesses a wide range of mathematical tasks. Items appear to assess deep math knowledge, math proficiency strands, and cognitively complex concepts.
- *Other considerations for use*: This assessment costs about $2 per student test, plus local scoring costs including scoring training (one or more sessions) and scoring time (10 to 20 student papers per scorer hour). For outside scoring and reporting services, costs are about $10 per student test. Annual membership fees are tiered based on the number of students served: $1,260 for entities serving fewer than 250 students to $31,500 for entities serving more than 100,000 students.

K6. **Star Math** is a web-based adaptive multiple-choice assessment for use in kindergarten to grade 12. No public information is available about the average number of items, and the average time to administer is 24 minutes.

- *Key publication:* White paper on Star Math, (Renaissance, 2020). Information about the assessment is available from the developer here.
- *Considerations for reliability:* No public information is available about reliability and validity for students in the priority communities. Additional testing is needed with students in the priority communities.
- *Relationship to math knowledge: Star Math* have a large math item bank. Items appear to assess deep math knowledge and cognitively complex concepts. Not enough information is available to determine if the measure assesses math proficiency strands.
- *Other considerations for use*: Cost is estimated at $4.95 per student for the 2020-2021 year. However, there is also a one-time fee of $1,599 per school for first-time users and an annual $750 per-school fee for hosting Renaissance software. Star Math are also available in Spanish.

K7. The **Iowa Algebra Aptitude Test** is a 50-minute exam for students in grades 7 and 8 and assesses students' readiness for Algebra I. It is aligned with the National Council of Teachers of Mathematics standards and is split into four parts: pre-algebraic number skills and concepts, interpreting mathematical information, representing relationships, and using symbols.

- *Key publication:* No publicly available evidence; documentation may be available through direct request to the measure developer. Information about the assessment is available from the developer here.
- *Considerations for reliability:* The developer asserts that measures are reliable and valid. No public information is available about reliability and validity for students in the priority communities. Additional testing is needed with students in the priority communities.
- *Relationship to math knowledge:* Items appear to assess deep math knowledge, math proficiency strands, and cognitively complex concepts.
- *Other considerations for use*: Cost is estimated to be about $5 per student, but it is not clear if other fees are associated with usage. Scoring and reporting services are available for an additional fee. Additionally, this assessment is designed for use in 7th and 8th grades, so it is not clear how useful it is for other grades.

K8. There are two different **Tests for Understanding Fractions**, one for 4th grade and one for 5th grade. The 4th-grade test includes 26 multiple-choice questions and uses items from the National Assessment of Educational Progress (NAEP) test, *Illustrative Mathematics,* and a fraction item bank. The 5th-grade test includes 18 multiple-choice questions and draws from items used in the NAEP test and the Partnership for Assessment of Readiness for College and Careers.

- *Key publication:* Jayanthi et al. (2017)
- *Considerations for reliability:* Jayanthi et al. (2017) reported reliability for samples of 4th-grade and 5th-grade students. The sample was about 30 percent Black, 15 percent Latino, 9 percent students with disabilities, and 11 percent English language learners. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with the priority communities.
- *Relationship to math knowledge:* Items assess deep math knowledge, math proficiency strands, and cognitively complex concepts.
- *Other considerations for use*: The instrument is available for free but requires permission from the authors. A Spanish version of the test is not currently available, but the panel believes that a translation would be easy to create. The developer (Instructional Research Group) can draw from a large normative sample to help with interpretation, by request.

Table 4. Math enjoyment constructs and structure

	E1. Attainment Value of Mathematics	E2. Beliefs, Engagement, and Attitude of Math Motivation Scale	E3. Expectancy-Cost-Value Scale	E4. Indiana Mathematics Belief Scale	E5. Math and Me Survey	E6. Math and Science Engagement Scales	E7. Math Identity Measures	E8. Mathematical Mindset	E9. Mathematics Attitude Inventory	E10. Mathematics Self-Efficacy Scale	E11. Sense of Connectedness to My Mathematics Classroom
Sense of belonging						S					F
Self-efficacy/confidence		I	S	S	S		I	I	S	F	
Enjoyment/general attitude		I			S	F	I	I	S		
Growth mindset				S				I			
Value and importance of math	F	I	S		S			I	S		

F = full measure needs to be used to assess construct; S = the instrument includes a subscale related to the construct; I = there are individual items related to the construct.

Measures of math enjoyment

Measures of math enjoyment in this section assess students' attitudes and mindsets about math. Table 4 provides information about the specific constructs included in each measure. Although some measures have individual items related to a specific construct, the panel strongly advises against pulling individual items from scales/subscales. It is unclear how individual items perform in isolation. If users extract individual items or subscales from instruments or combine them with their own items, they should psychometrically assess the combined instrument.

E1. The **Attainment Value of Mathematics** is an eight-item questionnaire on motivation in mathematics and positive math identity that uses a 6-point Likert scale. The measure has two subscales: importance of mathematics as self-defining and generalized value of mathematics. The measure is designed for grades 5 to 9. No information is available on time to administer.

- *Key publication:* Matthews (2018)
- *Considerations for reliability:* The instrument meets the required reliability threshold overall. All subscales meet the required reliability threshold based on prior research. Reliability was established with a sample that was 56 percent Black and 27 percent Latino students across schools in which 85 percent of students were eligible for free or reduced-price lunch. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with the priority communities.
- *Relationship to math enjoyment:* The full measure assesses the value and importance of math.
- *Other considerations for use*: This instrument can be implemented in an online setting and is available for free. It measures self-perceptions based on abilities that can change and grow over time. The scale was developed as part of an ongoing longitudinal study on motivation in mathematics.

E2. The **Beliefs, Engagement, and Attitude of Math Motivation Scale** is a 10-item survey that measures beliefs, engagement, and attitudes toward math on a dichotomous scale. This assessment is designed for grades 2 and 3. Time to administer is three minutes.

- *Key publication:* Orosco (2016)
- *Considerations for reliability:* The instrument meets the required reliability threshold overall. Item-level reliability also meets the required reliability threshold based on prior research. Reliability was established with a sample that was greater than 50 percent free or reduced-price lunch and 30 percent Latino. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with the priority communities. Prior research found the instrument to be psychometrically valid at the elementary level (grades 2 and 3). No evidence is available at the secondary level.
- *Relationship to math enjoyment:* Individual items assess self-efficacy/confidence, enjoyment/general attitude, and the value and importance of math.
- *Other considerations for use*: Measure is available without restrictions. Orosco (2016) included a sample of 2nd- and 3rd-grade students, so the instruments might need to be adapted for middle grades. The measure could be administered virtually.

E3. The **Expectancy-Cost-Value Scale** is a 10-item student survey instrument based on the expectancy-value theory of motivation that uses a 6-point Likert scale. The survey is designed to measure motivation in math and science in middle school students in grades 6 to 9. The instrument has three subscales: expectancy, cost, and value. The value scale includes utility value, attainment value, and interest/enjoyment value subscales and components. Time to administer is less than five minutes.

- *Key publications:* Kosovich et al. (2014), Lauermann et al. (2017), Simpkins et al. (2006), Wang (2012)
- *Considerations for reliability:* The instrument meets the required reliability threshold overall. All subscales meet the required reliability threshold based on prior research. This scale has been widely used and validated with many socioeconomically and racially diverse samples ranging from 3rd to 12th grade since the 1990s.
- *Relationship to math enjoyment:* The instrument includes subscales related to self-efficacy/ confidence and the value and importance of math.
- *Other considerations for use*: The instrument is available for use for free. It is a survey that could be administered virtually. It assesses the utility value, attainment value, interest/enjoyment value, and cost value for students (decision process weighing school work against other activities they could participate in, assessment of effort required for school work, and emotional cost).

E4. The **Indiana Mathematics Belief Scale** is a 35-item scale that students complete to measure their beliefs on constructs using a 5-point Likert scale. The measure assesses self-efficacy, identity, growth mindset, and related beliefs for secondary school and college levels. The instrument includes five subscales: effort, usefulness, difficult problems, understanding, and steps. The scale takes about 15 minutes to administer.

- *Key publications:* Kloosterman & Stage (1992), Ayebo & Mrutu (2019)
- *Considerations for reliability:* The instrument meets the required reliability threshold overall. Three of the five subscales meet the required reliability threshold. Two subscales (steps and difficult problems) do not meet the required threshold. Reliability was established with a sample of undergraduate college students, and no demographic information was presented. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with the priority communities and targeted age groups.
- *Relationship to math enjoyment:* The instrument includes subscales related to self-efficacy/ confidence and growth mindset.
- *Other considerations for use*: The instrument is available for free and could be administered virtually.

E5. The **Math and Me Survey** is a 27-item scale students in grades 3 to 6 complete to measure their beliefs on three constructs using a 5-point Likert scale. Subscales include mathematical self-perceptions, enjoyment of mathematics, and perceived usefulness of mathematics. No information is available on time to administer.

- *Key publication:* Adelson & McCoach (2011)
- *Considerations for reliability:* The instrument meets the required reliability threshold overall. All subscales meet the required reliability threshold based on prior research. Reliability was established with a sample of students in grades 3 to 6, including 34 percent non-White students. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with the priority communities.
- *Relationship to math enjoyment:* The instrument includes subscales related to self-efficacy/confidence, enjoyment/general attitude, and the value and importance of math.
- *Other considerations for use*: The instrument is available for free and could be administered virtually.

E6. The **Math and Science Engagement Scales** comprise of 33 items (17 positively worded and 16 negatively worded) that are on a 5-point Likert scale. Subscales include behavioral engagement, emotional engagement, cognitive engagement, and social engagement. The measure was tested on students in grade 5 to 12. The full scale takes about 7 to 10 minutes to administer.

- *Key publications:* Wang et al. (2016), Fredricks et al. (2016)
- *Considerations for reliability:* The instrument meets the required reliability threshold overall. All subscales meet the required reliability threshold based on prior research. Reliability was established with a sample of students in grades 6 to 12. This measure has been validated with racially/ethnically and socioeconomically diverse student samples through qualitative (focus group and cognitive interviewing) and quantitative methods (measurement invariance by race/ethnicity, socioeconomic status, and grade level) to support cultural responsiveness.
- *Relationship to math enjoyment:* The instrument includes a subscale related to sense of belonging. Overall, the instrument also assesses enjoyment/general attitude.
- *Other considerations for use*: The instrument is free and could be administered virtually. A shorter version is also available upon request. The survey has been used with English learners and translated into more than 10 languages. An audio version of questions is also available for student use. A teacher engagement scale is also available, which has 20 items on a 5-point Likert scale. Observational tools are also available.

E7. The **Math Identity Measures** is a 10-item subscale (from the Factors Influencing College Success in Mathematics Survey) that measures students' math identity and attitudes. Subscales include interest, recognition, and competence/performance. The measure was tested on students enrolled in college-level calculus classes across the United States. No information is available on time to administer.

- *Key publication:* Cribbs et al. (2015)
- *Considerations for reliability:* The instrument meets the required reliability threshold overall. All subscales meet the required reliability threshold based on prior research, except for the recognition subscales. The scale was used with priority communities, but reliability is not reported for racial or ethnic subgroups. Reliability was established in undergraduate students who were 7 percent Latino and 5 percent Black. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with targeted age groups and priority communities.
- *Relationship to math enjoyment:* The instrument includes items related to self-efficacy/confidence and enjoyment/general attitude.
- *Other considerations for use*: The instrument is free, but permission for use is required. It could be administered virtually.

E8. The **Mathematical Mindset** is a 29-item survey students complete to measure their beliefs and attitudes on math mindset, math as creative and connected, and fear of math on a 6-point Likert scale. The measure was designed for grades 6 to 8. The full scale takes about 7 to 10 minutes to administer.

- *Key publication:* Boaler et al. (2018)
- *Considerations for reliability:* Reliability information for the instrument overall is not available. One subscale (Fear of Math) meets the required reliability threshold, while the Mindset and Math as Creative and Connected subscales did not meet the required reliability threshold based on prior research. Analysis was conducted on a subset of nine items. Reliability was established with a sample of students in grades 6 to 8 that was 20 percent Latino, 2 percent Black, and 21 percent free and reduced-price lunch status students. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with the priority communities using the full survey items.
- *Relationship to math enjoyment:* The instrument includes items related to self-efficacy/confidence, enjoyment/general attitude, growth mindset, and the value and importance of math.
- *Other considerations for use*: This instrument has been used in a virtual setting.

E9. The **Mathematics Attitude Inventory** is a 48-item survey students complete to measure their beliefs and attitudes on six constructs: perception of mathematics teacher, anxiety toward mathematics, value of mathematics in society, self-concept in mathematics, enjoyment of mathematics, and motivation in mathematics. The measure was designed for students in grades 7 to 12. The measure takes about 20 minutes to administer.

- *Key publication:* Sandman (1980)
- *Considerations for reliability:* The instrument meets the required reliability threshold overall. All subscales meet the required reliability threshold based on prior research. Reliability was established with a sample in grades 7 to 12; no racial demographic information was provided. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with targeted grades and priority communities.
- *Relationship to math enjoyment:* The instrument includes subscales related to self-efficacy/confidence, enjoyment/general attitude, and the value and importance of math.
- *Other considerations for use*: This instrument is available for purchase, and it is unclear whether the publisher would allow use in a virtual setting.

E10. The **Mathematics Self-Efficacy Scale** is a nine-item survey students complete to evaluate their self-efficacy in the classroom and a test setting using a 5-point LIkert scale. The study examined students in grades 9 and 10. Administration should take about 10 to 15 minutes.

- *Key publication:* Nielsen & Moore (2003)
- *Considerations for reliability:* This instrument meets the required reliability threshold based on prior research. Reliability was established based on 300 high school students in Australia, with no information about sample demographics. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with targeted grades and priority communities.
- *Relationship to math enjoyment:* The instrument assesses self-efficacy/confidence.
- *Other considerations for use*: The measure is free and readily available.

E11. The **Sense of Connectedness to My Mathematics Classroom** is a nine-item survey using a 6-point Likert scale. It includes three subscales: Emotional Support from Math Teacher, My Math Class is Like a Family, and My Contributions are Valued. The study examined students in grades 5 to 9. No information is available on time to administer.

- *Key publication:* Maloney & Matthews (2020)
- *Considerations for reliability:* Reliability information for the instrument overall is not available. Two subscales (Emotional Support from Math Teacher and My Math Class is Like a Family) meet the required reliability threshold based on prior research. The subscale for My Contributions are Valued did not meet the required threshold. Reliability was established in a sample that was 56 percent Black and 30 percent Latino. Reliability is not reported for racial or ethnic subgroups. Additional testing is needed with targeted grades and priority communities.
- *Relationship to math enjoyment:* The instrument assesses sense of belonging. Survey responses are dependent on students' perceptions of the current classroom environment, rather than math more broadly.
- *Other considerations for use:* The instrument is available without restrictions on access. The means and standard deviations are provided for the subscales. It can be implemented in a virtual setting.

References

Adelson, J. L., & McCoach, D. B. (2011). Development and psychometric properties of the Math and Me survey: Measuring third through sixth graders' attitudes toward mathematics. *Measurement and Evaluation in Counseling and Development, 44*(4), 225–247. [E5]

Ayebo, A., & Mrutu, A. (2019). An exploration of calculus students' beliefs about mathematics. *International Electronic Journal of Mathematics Education, 14*(2), 385–392. [E4]

Boaler, J., Dieckmann, J. A., Pérez-Núñez, G., Sun, K. L., & Williams, C. (2018, April). Changing students' minds and achievement in mathematics: The impact of a free online student course. *Frontiers in Education, 3*, 26. [E8]

Cribbs, J. D., Hazari, Z., Sonnert, G., & Sadler, P. M. (2015). Establishing an explanatory model for mathematics identity. *Child Development, 86*, 1048–1062 [E7]

Fredricks, J. A., Wang, M. T., Schall, J., Hofkens, T. L., & Parr, A. (2016). Using qualitative methods to develop a survey measure of math and science engagement. *Learning and Instruction, 43*, 5–15. [E6]

Jayanthi, M., Gersten, R., Taylor, M. J., Smolkowski, K., & Dimino, J. (2017). *Impact of the Developing*

Mathematical Ideas professional development program on grade 4 students' and teachers' understanding of fractions (REL 2017–256). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. http://ies.ed.gov/ncee/edlabs [K4]

Kloosterman, P., & Stage, F. K. (1992). Measuring beliefs about mathematical problem solving. *School Science and Mathematics, 92*(3), 109–115. [E4]

Lauermann, F., Tsai, Y.-M., & Eccles, J. S. (2017). Math-related career aspirations and choices within Eccles et al.'s expectancy–value theory of achievement-related behaviors. *Developmental Psychology, 53*(8), 1540–1559. [E3]

Maloney, T., & Matthews, J.S. (2020). Teachers' critical care and students' feelings of connectedness in the urban mathematics classroom. *Journal of Research in Mathematics Education*. https://doi.org/10.1177/0042085919842625 [E11]

Matthews, J. S. (2018). When am I ever going to use this in the real world? Cognitive flexibility and urban adolescents' negotiation of the value of mathematics. *Journal of Educational Psychology, 110*(5), 726–746. https://doi.org/10.1037/edu0000242 [E1]

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. TIMSS & PIRLS International Study Center, Boston College.

National Research Council. (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: The National Academies Press. https://doi.org/10.17226/9822.

Nielsen, I. L., & Moore, K. A. (2003). Psychometric data on the Mathematics Self-Efficacy Scale. *Educational and Psychological Measurement, 63*(1), 128–138. [E10]

Orosco, M. J. (2016). Measuring elementary student's mathematics motivation: A validity study. *International Journal of Science and Mathematics Education, 14*(5), 945-958. [E2]

Renaissance Learning, Inc. (2020). *Research Foundation for Star Adaptive Assessments* [White paper]. http://doc.renlearn.com/KMNet/R001480701GCFBB9.pdf

Sandman, R. S. (1980). The Mathematics Attitude Inventory: Instrument and user's manual. *Journal for Research in Mathematics Education, 11*(2), 148–149. [E9]

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology, 42*(1), 70. [E3]

Wang, M. T. (2012). Educational and career interests in math: A longitudinal examination of the links between perceived classroom environment, motivational beliefs, and interests. *Developmental Psychology, 48*, 1643-1657. [E3]

Wang, M. T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Schall, J. (2016). The Math and Science Engagement Scale: Scale development, validation, and psychometric properties. *Learning and Instruction, 43*, 16–26. [E6]

Appendix A. Measure criteria

This set of criteria includes the three steps for assessing math measures. These criteria were used to assess measures in the menu, and can be applied to other measures as well. The menu of measures highlights measures that meet all or some of the criteria, and it documents which criteria each measure meets and does not meet. In addition to this set of general criteria, users assessing and selecting measures should also consider the context-specific questions listed in the menu. These questions will help assess whether a measure is a good fit in terms of content and feasibility.

The following high-level steps guide decision making when assessing whether a measure should be used for evaluation purposes:

- **Step 1.** Users should aim to use measures that meet all of the Step 1 criteria, which the panel identified as a minimal benchmark for measures to meet.
- **Steps 2 and 3.** Users should prioritize measures that meet some or all of the Step 2 and 3 criteria if possible. In schools that serve English learners, the measure should also meet criteria in Step 2a. Step 2a can be adapted for other languages as well. If no available measures meet criteria in Steps 2, 2a, and 3, then users could adapt measures as recommended in those steps.

Note that "priority communities" refers to the students whom the Middle Years Math grants seek to serve: students who are Black, Latino, and/or experiencing poverty.

Step 1: Determine if measures meet required criteria

- **Can be measured reliably across conditions**
  - *For standardized tests that have multiple items related to the same construct:* Internal consistency must be >0.70.
  - *For surveys with self-reported measures of beliefs, attitudes, and behaviors related to the same construct:* Internal consistency for subscales must be >0.70. All subscales must meet this minimum level.
  - *For measures that require subjective judgement to assign ratings or scores:* More than one rater should score a subset of assessments, and inter-rater reliability must be >0.70. Even if prior research demonstrates inter-rater reliability, grantees should establish inter-rater reliability with their own raters as well.
- **Measure is specific to math**

Step 2: Determine if measures meet the preferred criteria

- **Measure has demonstrated reliability and validity with priority communities**
  - Has demonstrated reliability and validity with students from priority communities OR has been qualitatively tested (through focus groups, cognitive interviewing, or talk-aloud protocols) with students from priority communities
  - Measure was modified to reflect findings from qualitative testing, and reliability and validity were reestablished on modified version

/ **Measure is available for use by programs without restrictions on access**

- Programs can access and use the measure without substantial burden.

/ **Guidance is available to help users consistently interpret results**

Linguistic accessibility criteria for users serving English learners

Step 2a: Determine if measures meet additional preferred criteria

/ **Measure is linguistically accessible**

- Directions and test items are presented in students' native language alongside original English version OR one or more tools is provided to support comprehension, including:
  - Audio versions of questions
  - Dictionaries
  - Pop-up glossaries
  - Audio glossaries
  - Printed glossaries
- Translations and tools include minimal barriers to use (e.g., minimal mouse clicking skills)
- Has been translated by a culturally competent translator with translation best practices, including:
  - Translator is a native speaker and considers differences in vocabulary, pronunciation, forms of speech, and idiomatic expressions in multiple dialects
  - Translator has math content knowledge (if applicable)
  - Translation performed by independent translators with translation reconciliation to resolve discrepancies
- Test items have been assessed for linguistic accessibility, including:
  - Equivalence of items in different languages assessed with differential item functioning
  - Technical terminology and translation assessed using cognitive interviews with students from priority communities to ensure that measure elicits intended mental processes
- Measure was modified to reflect findings from qualitative testing, and reliability and validity were re-established on modified version. ◢

Step 3: Determine if measures meet the following preferred content-specific criteria

For math knowledge measures

/ **Measure assesses deep knowledge of math, not rote memorization of math facts**

/ **Measure assesses one of the math proficiency strands (National Research Council, 2001)**

- Procedural fluency—ability to carry out procedures flexibly, accurately, efficiently, and appropriately
- Conceptual understanding—fundamental understanding of mathematical ideas, and the ability to integrate new ideas into understanding by connecting with previous knowledge
- Strategic competence—ability to formulate, represent, and solve mathematical problems
- Adaptive reasoning—capacity for logical thought, reflection, explanation, and justification
- Productive dispositions with regard to big ideas in mathematics

/ **Measure assesses cognitively complex concepts**

For math enjoyment measures

/ **Measure assesses one of the key math enjoyment constructs identified by the panel**

- Key constructs include sense of belonging, self-efficacy/confidence, enjoyment/general attitude, growth mindset, and value and importance of math.

/ **Measure is meant to assess state-based characteristics instead of trait-based characteristics**

- State-based characteristic refers to a temporary way of thinking, feeling, behaving, and so on, while a trait-based characteristic is considered to be a stable and enduring characteristic

This publication is based on research funded by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.