Secondary Writing Menu of Measures

Measuring the effects of writing instruction and interventions is one way to learn how to improve students' writing performance. Writing is foundational in daily life and academic achievement. Understanding effective writing instruction and intervention is critical to support writing development among students.

This menu describes a set of teacher- and student-level measures and approaches to understand the extent that writing instruction and interventions change how teachers think about and teach students about writing, and what students know and how they feel about writing. Sharing and administering high-quality measures will help the field understand how to improve students' engagement with and understanding of writing.

Expert panel members

Teacher measures

- **Linda Friedrich**, WestEd
- **Karen Harris**, Arizona State University
- **Troy Hicks**, Central Michigan University
- **Nicole Merino**, Stanford University
- **Eugenia Mora-Flores**, University of Southern California, Rossier School of Education
- **Detra Price-Dennis**, Columbia University
- **Kay Wijekumar**, Texas A&M University

Student measures

- **Steve Graham**, Arizona State University
- **Ruth Wei**, Envision Learning Partners
- **Eugenia Mora-Flores**, University of Southern California, Rossier School of Education
- **Joshua Wilson**, University of Delaware
- **Maisha T. Winn**, University of California Davis

The authors originally developed this menu of measures, in consultation with a panel of experts, as a guide for the Bill & Melinda Gates Foundation's Secondary Writing grantees who sought to innovate and test automated writing feedback tools. Automated writing feedback tools (also sometimes referred to as automated writing evaluation tools) are digital writing tools that use artificial intelligence or machine learning technology to provide formative feedback on student essays or sentences as students write. The goal of the Secondary Writing grant portfolio was to develop, refine, and scale evidence-based solutions (programs, products, or practices) that enable students in grades 6 through 12 who are Black, Latino, and/or experiencing poverty (the communities in focus for the grants) to be engaged in argumentative writing and be on track with college- and career-level competencies. Grantees, in collaboration with their evaluation technical assistance providers, used the menu to select high-quality, common measures for their measurement and evaluation work. This menu of measures has been adapted for broader use.

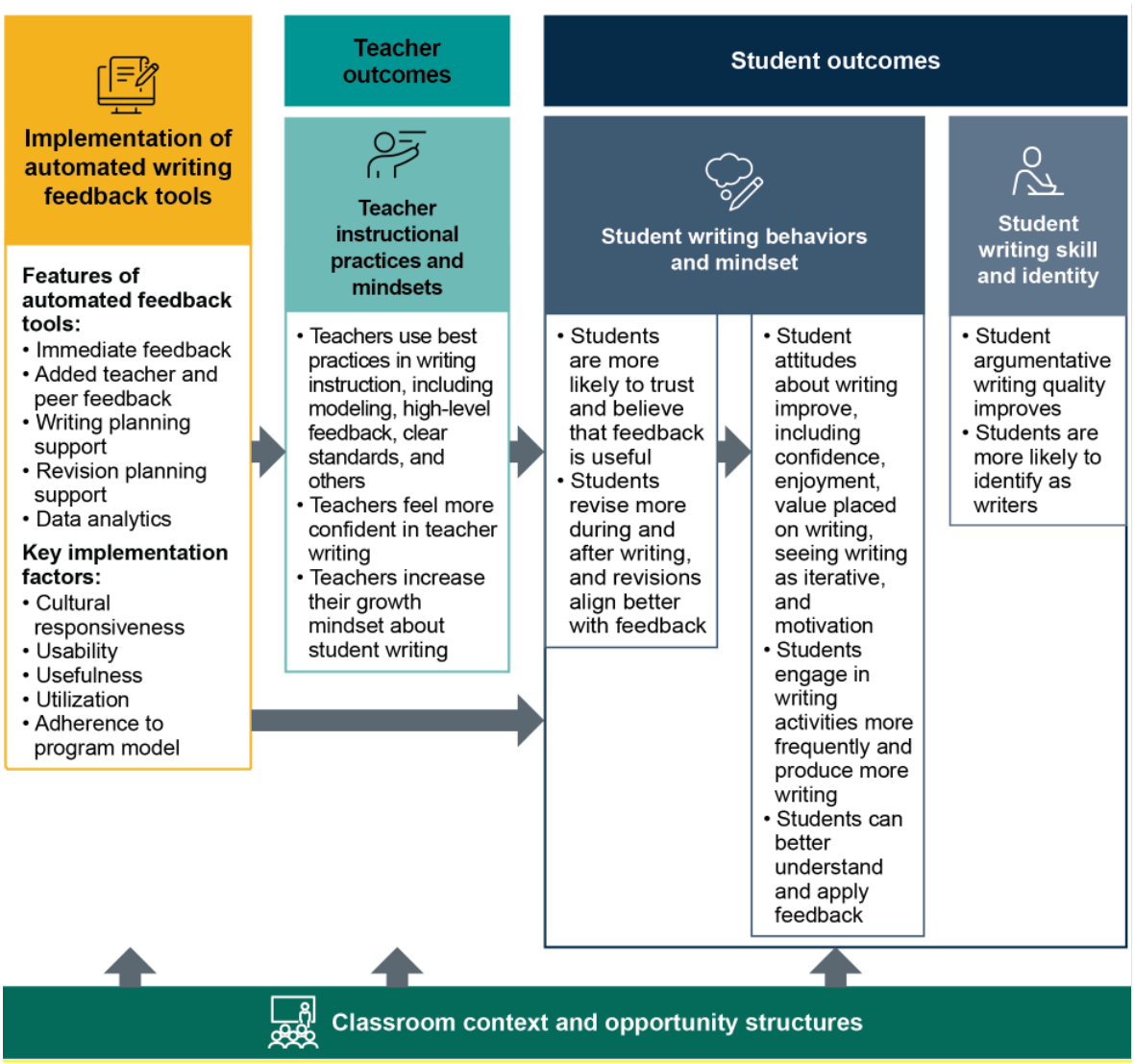
Specifically, this menu of measures is designed to help districts, researchers, funders, and organizations implementing programs learn about how to measure teachers' writing instructional practices and mindsets; students' writing behaviors and mindsets; and students' argumentative writing skills in a valid, reliable, and accessible way. When used at multiple points over time, these measures can provide information about how a particular approach to teaching or

supporting learning of writing affects students and teachers. The menu is not a comprehensive list of every writing teaching and learning measure available in the field. It is a curated list of measures developed in partnership with panels of researchers in the field of writing instruction and measurement with a focus on grades 6 to 12 and argumentative writing. The menu includes measures that meet some or all of a set of preferred criteria developed by the panel (see Appendices A and B).

One goal of writing instruction and interventions is to improve the quality of students' argumentative writing and how likely they are to identify as writers. Argumentative writing requires writers to take a stance on a topic and explain and persuade their audience of their position, typically through supporting evidence and addressing counterclaims and counterarguments. Argumentative writing is believed to help students develop critical thinking and persuasive skills and enhance broad thinking (Ray et al. 2018; MacArthur et al. 2015). Argumentative writing is a key component of the College and Career Readiness Anchor Standards for Writing defined by the Common Core State Standards for students in grades 6 through 12 (Common Core Standards Initiative, 2010).

Figure 1 shows a conceptual framework of how implementation of automated writing feedback tools, teacher instructional practices and mindsets, and classroom context and opportunity structures could affect different aspects of students' writing behaviors and mindsets, and ultimately their writing skill and identity. Students' writing behaviors and writing mindsets are shorter-term student outcomes that affect longer-term outcomes such as understanding and application of feedback, writing identity, and improvements in features of argumentative writing. Writing behaviors and mindsets do not on their own indicate that writing will be high quality, but they have been shown to be important factors in writing development.

Figure 1. Conceptual framework for improving writing skills and identity



This menu recommends instruments or approaches for measuring each of five key areas: teacher writing instructional practices, teacher writing mindsets, student writing behaviors, student writing mindsets, and student argumentative writing skills. For each instrument or approach, we provide a brief description, key publications, considerations for reliability, relationship to writing performance, and other considerations for use.

Although not covered in depth in this menu, measuring what happens in the classroom is important to understand the structures that enable or constrain students. See Box 1 for observation instruments and surveys that could be used to measure classroom context and opportunity structures.

Box 1. Measuring classroom context and opportunity structures

Opportunity structures are aspects of the school environment—interpersonal, instructional, and institutional factors—that impact a students' sense of belonging (Gray et al., 2018). Interpersonal opportunity structures refer to the social ties and connections between students and their educators and peers. Instructional opportunity structures refer to cultural alignment between the classroom setting, including educator led activities, and that of the student in a way that upholds the student's esteemed cultural meaning systems. Institutional opportunity structures refer to the process of eliminating structural barriers that devalue the experiences the experiences of specific groups of students (such as Black or Latino students or students experiencing poverty) in the school and community. Each of these structures play an important role in shaping student classroom experiences, including their sense of agency, their motivation, and their learning opportunities.

The student outcomes covered in this menu have the potential to change over time if teachers provide inclusive and supportive opportunity structures for students. Since these outcomes are direct consequences of how students are taught and supported in classrooms, it can be beneficial to pair the use of student outcome measures with measures of classroom context and opportunity structures.

Panelists recommended the following classroom observation instruments and surveys as reliable measures of classroom context and opportunity structures:

- Copilot-Elevate Student Survey is an 18-item student survey used to measure the quality of student learning conditions. The Project for Education Research that Scales (PERTS) developed it as part of a professional development platform.
- 5Essentials Survey is a survey used to measure aspects of school climate associated with improved student outcomes in elementary and high schools. The instrument was developed by the University of Chicago Consortium on School Research.

Overview of measures

The menu includes measures in two areas of teacher outcomes and three areas of student outcomes:

1. **Teacher writing instructional practices.** Teachers' use of promising practices in writing instruction (including modeling, high-level feedback, clear standards, and others)
2. **Teacher writing mindsets.** Teachers' beliefs about themselves, beliefs they hold about their students, and beliefs about teaching
3. **Student writing behaviors.** Students' frequency of writing in school and out of school, quantity of writing output, and plans for writing
4. **Student writing mindsets.** Students' confidence in their writing, how much they enjoy and value writing, how motivated they feel to write, how possible they believe it is to improve their writing, and their openness to feedback
5. **Student argumentative writing skills.** Quality of students' argumentative writing

Teacher measures

Table 1. Teacher measures by outcome area

Outcome area	Measures
Teacher writing instructional skills and practices	• Teachers' Use of Evidence-Based Writing Practices Scale [a] • College-Ready Writers Program Classroom Log • Protocol for Language Arts Teaching Observation
Teacher writing mindsets	• Teacher Efficacy Scale [a] • Teachers' Sense of Self-Efficacy Scale • Implicit Theories of Intelligence Scale • Teacher Attitudes Toward Writing Scale [a] • Writing Orientation Scale [a]

[a] Subscale of the National Survey of Teachers' Preparation and Practices in Teaching Writing.

Student measures

Table 2. Student measures by outcome area

Outcome area	Measures
Student writing behaviors	• Quantity of writing output • Writing Activities and Motivation Scale • 5-point scale for rating student plans • Time spent on writing
Student writing mindsets	• Self-Efficacy for Writing Scale • Implicit Theories of Writing Scale • Self-Beliefs, Writing-Beliefs, and Attitude Survey • Writing Disposition Scale

Outcome area	Measures
	• Liking Writing Scale • Writing Attitudes Survey • Writing Achievement Goals Scale • Beliefs about Writing Survey • Writing Activities and Motivation Scale • Writing Motivation and Engagement Scale • Writing Motivation Scale
Student argumentative writing skills	• Smarter Balanced Argumentative Performance Task Writing Rubric (Grades 6–11) • Advanced Placement (AP) English Language and Composition Scoring Rubric • PARCC/New Meridian • Literacy Design Collaborative Student Work Rubric for Argumentation Tasks • Score Basic Elements

Assessment of measures

This section summarizes the factors that Mathematica and panelists considered in identifying teacher and student measures, the criteria established by each panel to assess the measures, and the expert panels' assessments of the quality of teacher and student measures.[1]

Users should consider these factors when selecting and using both teacher and student measures:

- **Reliability.** What is known about whether scores are consistent across individuals, context, time, items, or raters? Prior research might show evidence of reliability, but users should verify that this reliability data is based upon assessments conducted with sample populations that are similar to the one they intend to assess (such as a similar percentage of students of color).
- **Validity.** Do the scores represent what they intend to measure?
- **Feasibility to administer.** Are instruments burdensome to administer or score, and are the barriers in terms of cost and licensing minimal? Can the instruments be administered virtually?

Several measures included in this menu are designed to be collected from students. For these types of measures, users should also consider the following:

- **Cultural responsiveness.** Is the measure reliable and valid for use in classrooms that include students from the communities in focus? For instruments that students complete, are the items relevant for the sociocultural values and experiences of students from the communities in focus?
- **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?

Within these factors, the specific criteria used to assess each measure differed slightly for teacher and student measures (see Appendices A and B, respectively).

Tables 3 and 4 show which assessment criteria each measure meets (and does not meet). 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 met the criterion of interest. The information used to assess measures in this menu was documented in 2020, and updated information on the measures summarized in the tables might have been published since then. When using this resource, users should also

[1] The measures described in this section were identified and assessed by the expert panel. Expert panelists first recommended existing measures in the three outcome areas covered in this menu: (1) writing behaviors, (2) writing beliefs and attitudes, and (3) argumentative writing. The expert panel then assessed each recommended measure against the criteria described in Appendices A and B.

consider a set of context-specific questions to determine whether the recommended measures are a good fit for the local context (see Box 2).

Box 2. Context-specific questions

Consider context-specific questions to determine fit of measures

Questions to assess validity:

- Is the measure designed to capture an outcome in the intervention's theory of change?
- Is the measure aligned with an outcome expected to change at the point of implementation when it is planned to be used? Longer-term outcomes should not be measured until the time at which the intervention would be expected to produce improvements.
- Is there evidence that the measure is predictive of expected longer-term outcomes as defined by the intervention's theory of change?
- Does the design of the measure match the intended use:
  - Is the measure designed to be a formative assessment, that is, intended to be used during a unit or course to measure progress and learning? Or is it designed to be a summative assessment, that is, intended to measure what students have learned at a defined end point of a unit or course?
  - 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 Spanish 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 the measure feasible to administer given the local context of the school or classroom where it would be administered?

Table 3. Teacher measures: Assessment criteria met and not met

	Teacher writing instructional skills and practices			Teacher writing mindsets				
	Teachers' Use of Evidence-Based Writing Practices Scale[a]	College-Ready Writers Program Classroom Log	Protocol for Language Arts Teaching Observation	Teacher Efficacy Scale[a]	Teacher Sense of Self-Efficacy Scale	Implicit Theories of Intelligence Scale	Teacher Attitudes Towards Writing Scale[a]	Writing Orientation Scale[a]
Measure type TS = teacher survey, TL = teacher log, O = observation	TS	TL	O	TS	TS	TS	TS	TS
Measure has adequate evidence of score reliability (such as internal consistency, inter-rater reliability)[b]	Yes	Yes	Yes[c]	Yes	Yes[d]	Yes	Yes	Yes[e]
Reliability has been established with teachers of secondary school students	No	Yes	Yes	No	No	Yes	No	No
Measure has clearly defined topics or constructs measured by each subscale	Yes	No	Yes	Yes	Yes	n.a.	Yes	Yes
Measure is available for use without restrictions on access	Yes	No	No	Yes	Yes	Yes	Yes	Yes
Measure is specific to writing	Yes	Yes	No	Yes	No	No	Yes	Yes
Measure was developed for research purposes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Measure includes clear instructions for ease of use (applies to teacher logs only)	n.a.	No	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Measure gauges frequency and intensity of focal instructional activities	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Measure is sensitive to change with intervention	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Materials are available for training raters and ensuring that they apply the measure in the same way	n.a.	n.a.	Yes	n.a.	n.a.	n.a.	n.a.	n.a.
Measure is culturally responsive for use with students from communities in focus	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

[a] Subscale of the National Survey of Teachers' Preparation and Practices in Teaching Writing.

[b] The panel recommended setting a threshold of 0.70 for internal consistency of scales and 0.70 for inter-rater reliability of observational measures.

[c] Reliability established using an earlier version of the instrument.

[d] The Teacher Sense of Self-Efficacy Scale includes a 24-item long form, a 12-item short form, and an alternate form focused on language instruction. We do not have reliability information for the alternate form.

[e] There is sufficient reliability for the correct writing subscale (internal consistency is 0.70). Reliability for the explicit instruction subscale is 0.64 afnd is 0.60 for the natural learning subscale. All measures in this table can be adapted for remote administration.

n.a. = not applicable.

Table 4. Student measures: Assessment criteria met and not met

	Student writing behaviors				Student writing mindsets											Student argumentative writing skills				
	Quantity of writing output	Writing Activities and Motivation Scale	5-point scale for rating student plans	Time spent on writing	Self-Efficacy for Writing Scale	Implicit Theories of Writing Scale	Self-Beliefs, Writing-Beliefs, and Attitude Survey	Writing Disposition Scale	Liking Writing Scale	Writing Attitudes Survey	Writing Achievement Goals Scale	Beliefs about Writing Survey	Writing Activities and Motivation Scale	Writing Motivation and Engagement Scale	Writing Motivation Scale	Smarter Balanced Argumentative Performance Task Writing Rubric	AP English Language and Composition Scoring Rubric	PARCC/New Meridian	Literacy Design Collaborative Student Work Rubric for	Score Basic Elements
Construct F = frequency of writing occasions in school; Q = quantity of writing output; P = plans for writing; V = perceived value of writing; C = confidence in writing; E = enjoyment of writing; M = motivation to write; O = Overall quality of argumentative writing	Q	F	P	F	C	C	C	C	E	E	V	V	M	M	M	O	O	O	O	O
Criterion																				
Measure is culturally responsive for use with students from communities in focus	n.a.	No	No	n.a.	No	No	No	No	No	No	No	Yes	No	No	Yes	Yes	No	No	Yes	No
Measure has adequate evidence of score reliability (such as internal consistency, inter-rater reliability)[a]	n.a.	Yes	Yes	n.a.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes[b]	Yes	Yes	Yes

	Student writing behaviors				Student writing mindsets											Student argumentative writing skills				
	Quantity of writing output	Writing Activities and Motivation Scale	5-point scale for rating student plans	Time spent on writing	Self-Efficacy for Writing Scale	Implicit Theories of Writing Scale	Self-Beliefs, Writing-Beliefs, and Attitude Survey	Writing Disposition Scale	Liking Writing Scale	Writing Attitudes Survey	Writing Achievement Goals Scale	Beliefs about Writing Survey	Writing Activities and Motivation Scale	Writing Motivation and Engagement Scale	Writing Motivation Scale	Smarter Balanced Argumentative Performance Task Writing Rubric	AP English Language and Composition Scoring Rubric	PARCC/New Meridian	Literacy Design Collaborative Student Work Rubric for	Score Basic Elements
Reliability has been established with secondary school students	n.a.	Yes	No	n.a.	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes[b]	Yes	Yes	Yes
Measure has clearly defined topics or constructs measured by each subscale	n.a.	Yes	No	n.a.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Measure is available for use without restrictions on access	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Measure is specific to writing	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Guidance is available to consistently interpret results	No	Yes	No	No	No	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No
Measure uses student-friendly accessible language	n.a.	Yes	n.a.	n.a.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	n.a.	n.a.	n.a.	n.a.	n.a.
Measure is linguistically accessible to English learners who also speak Spanish	n.a.	No	n.a.	n.a.	Yes	No	No	No	No	No	No	No	No	Yes	Yes	n.a.	n.a.	n.a.	n.a.	n.a.

[a] The panel recommended setting a threshold of 0.70 for internal consistency of scales and 0.70 for inter-rater reliability of observational measures.

[b] Score reliability for the AP English Language and Composition Scoring Rubric is based on the 1999 version of the rubric, the AP English Language 1999 Scoring Guidelines.

n.a. = not applicable

Description of measures

This section describes each measure in the menu, including additional considerations for implementation and interpretation. The information in this section can help users select a measure related to their theory of change.

Teacher measures

Measures of teacher writing instructional skills and practices

T1. The **Teachers' Use of Evidence-Based Writing Practices Scale** is a subscale on the National Survey of Teachers' Preparation and Practices in Teaching Writing. There are two versions of the subscale, an 18-item version used by Brindle et al. (2016) and a 15-item subscale used by Graham et al. (2014). Teachers respond to items using an 8-point scale indicating how frequently they used a practice. Each item corresponds to an evidence-based practice, including providing written feedback on students' essays and establishing specific goals for students' writing, teaching students to self-regulate the writing process.

- Key publications: Brindle et al. (2016), Graham et al. (2014), MI Write study summary (2023), Ecree study summary (2023)
- Considerations for reliability:
  - Brindle et al. (2016) found that the subscales meet the required reliability threshold (internal consistency of evidence-based teaching was 0.90; evidence-based writing was 0.80). Reliability was established with 3rd- and 4th-grade teachers. A similar version was used at the middle school level (Graham et al., 2014), but reliability data for this version are not published. The measure's reliability may perform differently with samples of students from secondary grade levels.
  - In a study of the automated writing feedback tool MI Write, the 15-item scale met the required reliability threshold (internal consistency was 0.94 at baseline and 0.93 at follow-up). The study took place during the 2021–2022 school year and included 7th- and 8th-grade English language arts (ELA) teachers from three school districts serving high proportions of students who are Black, Latino, and/or experiencing poverty (60 percent of students who participated in the study were Black and/or Latino, and 60 percent were eligible for free or reduced-price lunch).
  - In a study of the automated writing feedback tool Ecree, the 15-item scale met the requirement reliability threshold (internal consistency was 0.94 at baseline and 0.95 at follow-up). The study took place during the 2021–2022 school year and included grade 8–11 ELA teachers from two school districts (22 percent of students who participated in the study were Black and/or Latino).
- Relationship to writing performance: A series of meta-analyses assessed the relationship between each evidence-based practice and improved student writing. Average effect sizes for items across this subscale ranged from 0.24 to 1.17 (Graham et al., 2011, 2012).

- Other considerations for use: Brindle et al. (2016) used this survey with 3rd- and 4th-grade teachers. Graham et al. (2014) used a similar version with middle school teachers but did not report reliability data. The full survey instrument also includes scales on pre-service and in-service preparation and on teacher beliefs.

T2. The **College-Ready Writers Program Classroom Log** is a tool for teachers to record the frequency, intensity, and specific elements of writing instruction. Teachers record the time spent writing, length of writing assigned, purposes of writing that day, relative emphasis on facets and genres of writing, perceptions of self-efficacy, and professional development experiences. The log assesses practices that are specific to argumentative writing instruction, rather than more general language arts instruments.

- Key publication: Gallagher et al. (2017)
- Considerations for reliability: Gallagher et al. (2012) found that an adapted version of the log met the required reliability threshold (inter-rater reliability between teachers and researchers was 0.80 overall, with a range of 0.61 to 0.96 across measures). Reliability was established with 7th- and 8th-grade teachers. Teachers were instructed to complete the log at the end of each class period or later in the day so they could reliably recall their activities. Reliability for the version Gallagher et al. (2017) used is not published but is available from the authors. The panel is not aware of training materials available for users.
- Relationship to writing performance: Gallagher et al. (2017) did not report information about validity.
- Other considerations for use: In a research setting, Gallagher et al. (2017) administered the log by having teachers complete it daily over a two-week period in the fall and a two-week period in the spring.

T3. The **Protocol for Language Arts Teaching Observation (PLATO)** is a rubric used to score 13 elements of ELA instruction on a scale of 1 to 4. The rubric covers four domains: Disciplinary Demand of Classroom Talk and Activity, Contextualizing and Representing Content, Instructional Scaffolding, and Classroom Environment. PLATO is designed for research purposes rather than diagnostics or evaluation. Each domain can be used separately.

- Key publications: PLATO Version 5.0 (Stanford University, 2013), Cor (2011)
- Considerations for reliability: Cor (2011) found that a prior version of the PLATO rubric met the required reliability threshold (internal consistency was 0.81 for the entire rubric and greater than 0.70 for each domain). Additional information is needed about reliability and validity of the most recent version (5.0). A time-intensive training is available for observers, and guidance indicate that inter-rater reliability should be established within specific groups of observers prior to administration.
- Relationship to writing performance: Cor (2011) did not report information about validity.
- Other consideration for use: This instrument contains 13 elements of ELA instruction: Purpose, intellectual challenge, representation of content, connections to prior knowledge, connections to personal and cultural experience, modeling and use of models, strategy use

and instruction, feedback, classroom discourse, text-based instruction, accommodations for language learning, behavior management, and time management.

Measures of teacher writing mindsets

T4. The **Teacher Efficacy Scale** is a nine-item scale on the National Survey of Teachers' Preparation and Practices in Teaching Writing. Teachers respond to each item using a 6-point Likert-type scale. Items cover teacher beliefs about their effectiveness to teach writing generally, to teach specific aspects of writing, and to address specific student issues. This scale can be used independently from other scales on the National Survey of Teachers' Preparation and Practices in Teaching Writing.

- Key publications: Brindle et al. (2016), MI Write study summary (2023), Ecree study summary (2023)
- Reliability:
  - Brindle et al. (2016) found that this scale met the required reliability threshold (internal consistency was 0.89). Reliability was established with 3rd- and 4th-grade teachers. The measure's reliability may perform differently with samples of teachers from secondary grade levels.
  - In a study of the automated writing feedback tool MI Write, the nine-item scale met the required reliability threshold (internal consistency was 0.79 at baseline and 0.83 at follow-up). The study took place during the 2021–2022 school year and included 7th- and 8th-grade ELA teachers from three school districts serving high proportions of students who are Black, Latino, and/or experiencing poverty (60 percent of students who participated in the study were Black and/or Latino, and 60 percent were eligible for free or reduced-price lunch).
  - In a study of the automated writing feedback tool Ecree, the nine-item scale met the required reliability threshold at baseline (internal consistency was 0.79) but not at follow-up (internal consistency was 0.55). The study took place during the 2021–2022 school year and included grade 8–11 ELA teachers from two school districts (22 percent of students who participated in the study were Black and/or Latino).
- Relationship to writing performance: Brindle et al. (2016) did not report information about validity.
- Other considerations for use: The full National Survey of Teachers' Preparation and Practices in Teaching Writing includes scales on pre-service and in-service preparation, instructional practices, and attitudes toward teaching and writing.

T5. The **Teachers' Sense of Self-Efficacy Scale** is a 24-item instrument to measure teacher efficacy. A short-form version with 12 items is also available. The scale was developed by researchers at Ohio State University. The three subscales on the long and short forms can be used separately. The instrument includes three subscales: efficacy for instructional strategies, efficacy for student engagement, and efficacy for classroom management. The version that focuses on language instruction includes specific items about efficacy in writing instruction.

- Key publication: Tschannen-Moran and Hoy (2001)
- Considerations for reliability: Tschannen-Moran and Hoy (2001) found that both the long- and short-form versions met the required reliability threshold (internal consistency was 0.94 and 0.90, respectively). Reliability has not been established with secondary school teachers. The measure's reliability may perform differently with samples of teachers from secondary grade levels. Reliability was not published for the alternate version focused on language instruction.
- Relationship to writing performance: Tschannen-Moran and Hoy (2001) found that both the long- and short-form versions had sufficient construct validity when compared to other measures of teacher efficacy. The strongest correlations are with scales that measure personal teaching efficacy.
- Other considerations for use: Panelists believe that measures of attitudes will only be sensitive to change with intensive professional development.

T6. The **Implicit Theories of Intelligence Scale** is a six-item scale about teachers' beliefs related to the concept of growth mindset. Teachers respond to each item using a 6-point Likert-type scale. The original scale was developed by Dweck and Henderson (1989). A version for teachers' beliefs about growth mindset in students is available in Looney (2003).

- Key publication: Looney (2003)
- Considerations for reliability: Looney (2003) found that this scale met the required reliability threshold (internal consistency was 0.83). Prior adaptations made for the student scale were found to be reliable, but additional testing is needed for a teacher version specific to writing.
- Relationship to writing performance: Looney (2003) did not report information about validity.
- Other considerations for use: Panelists have some concerns about ceiling effects for this measure. Users should examine baseline scores to determine if there is room for improvement. Also, the scale relates to teacher growth mindset about students. It is not specific to writing. Other researchers have adapted the scale for measuring student growth mindset about writing; similar adaptations could be made for teachers.

T7. The **Teacher Attitudes Toward Writing Scale** is from the National Survey of Teachers' Preparation and Practices in Teaching Writing. The survey includes a six-item scale on teacher attitudes toward writing, plus one additional item on teacher attitudes toward teaching writing. Teachers respond to each item using a 6-point Likert-type scale. Items cover teachers' beliefs about whether they are good writers, write for different purposes, enjoy writing and learning to write, and enjoy teaching writing. This six-item scale can be used independently from other scales on the National Survey of Teachers' Preparation and Practices in Teaching Writing.

- Key publication: Brindle et al. (2016)
- Considerations for reliability: Brindle et al. (2016) found that this scale met the required reliability threshold (internal consistency was 0.87). Reliability was established with 3rd- and 4th-grade teachers. The measure's reliability may perform differently with samples of teachers from secondary grade levels.

- Relationship to writing performance: Brindle et al. (2016) did not report information about validity.
- Other considerations for use: Brindle et al. (2016) used this survey with 3rd- and 4th-grade teachers; adaptations might be needed for secondary grades. The full survey includes scales on pre-service and in-service preparation, instructional practices, and self-efficacy in teaching.

T8. The **Writing Orientation Scale** from the National Survey of Teachers' Preparation and Practices in Teaching Writing includes 13 items on teacher beliefs about how writing should be taught. Teachers respond to each item using a 6-point Likert-type scale. Items cover teachers' beliefs about the role of correct writing, explicit instruction, and natural learning. The scale was originally developed by Graham and colleagues in 2001 (Graham et al., 2001). This scale can be used independently from other scales on the National Survey of Teachers' Preparation and Practices in Teaching Writing.

- Key publication: Brindle et al. (2016), Graham et al. (2001)
- Considerations for reliability: Brindle et al. (2016) found that this scale met the required reliability threshold for the correct writing subscale (internal consistency was 0.70). Reliability for the explicit instruction subscale was 0.64 and was 0.60 for the natural learning subscale. Reliability was established with 3rd- and 4th-grade teachers. The measure's reliability may perform differently with samples of teachers from secondary grade levels.
- Relationship to writing performance: Brindle et al. (2016) did not report information about validity.
- Other considerations for use: Brindle et al. (2016) used this survey with 3rd- and 4th-grade teachers; adaptations might be needed for secondary grades. The full survey includes scales on pre-service and in-service preparation, instructional practices, and self-efficacy in teaching.

Student measures

Measures of student writing behaviors

Quantity of writing output

S1. The **quantity of writing output** in a digital writing tool, such as a word processor or automated writing feedback tool, is the number of words, sentences, and paragraphs the student writes. This measure can be tracked automatically in many digital writing tools.

- Key publication: Morphy and Graham (2012)
- Considerations for reliability: Not applicable.
- Relationship to writing performance: Morphy and Graham's (2012) meta-analysis of 13 studies found measures of writing quantity (for example, total number of words written, total sentences written, total number of idea units) to be good predictors of writing quality.

- Considerations for use: Writing output can be tracked using digital tools like Microsoft Word, GoogleDocs, or another digital writing tool. It can also be calculated manually if the text is handwritten, in which case a typical definition of a word is one or more letters separated from other letters by a space.

S2. The **Writing Activities and Motivation Scale** is a student survey with two main sections: (1) 30 items related to writing motivation on an 11-point scale ranging from totally disagree to totally agree and (2) 10 items related to frequency of writing activities on a 5-point scale ranging from almost never to almost daily. The writing activity portion asks students to self-report the frequency of various writing activities in and out of school. This measure is also listed in the section on measures of student writing mindsets.

- Key publication: Troia et al. (2013)
- Considerations for reliability: Troia et al. (2013) found that the required reliability threshold was met for the motivational beliefs scale (internal consistency was 0.88) but not for the achievement goal orientation scale. The study included 618 students in grades 4 to 7 and grades 9 and 10 (23 percent of students who participated in the study were Black and/or Latino); about one-third of the participating students were in grades 9 and 10.
- Relationship to writing performance: Troia et al. (2013) found that motivational beliefs were positively related to narrative writing performance.
- Other considerations for use: The survey items are not published in the article but are available from the authors.

Plans for writing

S3. The **5-point scale for rating student plans** is a scoring system for student writing plans. For example, plans receive a score of 0 if no plan is drafted, 1 if the plan is an exact copy of the composition, or 4 if the students used a sophisticated planning strategy such as a web, outline, or genre-specific planning strategy.

- Key publication: Wijekumar et al. (2019)
- Considerations for reliability: Wijekumar et al.'s (2019) study found that the 5-point scale met the required reliability threshold (inter-rater reliability was 0.88). The study included 179 5th-grade students (46 percent of students who participated in the study were Black and/or Latino).
- Relationship to writing performance: Wijekumar et al. (2019) found that complexity of writing plans was related to writing performance.
- Other considerations for use: The study includes a brief description of the rating levels but does not provide a rubric or training materials. The rubric requires manual rating by teachers, which would require training and establishing inter-rater reliability; it also could be time intensive.

Frequency of writing occasions in school

S4. The **time spent on writing** is the number of minutes the student spends engaged in writing. This measure can be tracked automatically in many digital writing tools.

- Key publication: Graham et al. (2012)
- Considerations for reliability: Not applicable.
- Relationship to writing performance: Graham et al.'s (2012) meta-analysis of five studies found that increasing the time students wrote had a corresponding increase in writing quality at the elementary level (grades 2–6). No evidence is available at the secondary level.
- Other considerations for use: Amount of time is likely to be tracked similarly across digital writing tools. Users should consider whether time tracking requires students to manually log in and log out, which could threaten reliability.

Measures of student writing mindsets

Confidence in writing

S5. The **Self-Efficacy for Writing Scale** is a 16-item student survey measuring ideation, conventions, and self-regulation using a 100-point scale ranging from no confidence to complete confidence.

- Key publications: Bruning et al. (2013), MI Write study summary (2023), Ecree study summary (2023)
- Considerations for reliability:
  - Bruning et al. (2013) conducted two studies, one with 697 middle schoolers and one with 563 high schoolers. Both studies found that all three subscales (ideation, conventions, and self-regulation) met the required reliability threshold (internal consistency ranged from 0.85 to 0.92 across subscales and studies). In both studies, most students were White (15 to 18 percent of each sample was Black or Hispanic).
  - In a study of the automated writing feedback tool MI Write, the overall scale and each of the three subscales met the required reliability threshold (internal consistency at baseline and follow-up was 0.95 for the overall scale and ranged from 0.88 to 0.93 for the subscales). The measures also met the required reliability threshold for the subsample of students who are Black, Latino, and/or eligible for free or reduced-price lunch. The study took place during the 2021–2022 school year and included 7th- and 8th-grade ELA teachers from three school districts serving high proportions of students who are Black, Latino, and/or experiencing poverty (60 percent of students who participated in the study were Black and/or Latino, and 60 percent were eligible for free or reduced-price lunch).
  - In a study of the automated writing feedback tool Ecree, two of the subscales were administered: ideation and convention. Both subscales met the required reliability threshold (internal consistency at baseline and follow-up ranged from 0.89 to 0.91). The

measures also met the required reliability threshold for the subsample of students who were Black and/or Latino. The study took place during the 2021–2022 school year and included grade 8–11 ELA teachers from two school districts (about 22 percent of students who participated in the study were Black and/or Latino).

- Relationship to writing performance: Bruning et al. (2013) found all three subscales were related to self-reported writing grades as well as to performance on statewide assessments. The relationships between the ideation subscale and test performance and between the self-regulation subscale and test performance were lower than the relationship for the conventions subscale.
- Other considerations for use: In Bruning et al. (2013), students completed a paper-based version of the survey in approximately 20 minutes. There is little information available about the extent to which student perceptions of their writing self-efficacy normally change over time. However, in two studies of automated writing feedback tools, MI Write and Ecree, mean scores ranged from 52 to 74 (depending on the scale or subscale and study sample) at the start of the school year among students receiving typical ELA instruction. By spring data collection, mean scores increased on average by around 1 to 3.5 points, with growth around 1 point in the Ecree study (typically about four to five months of growth) and around 2.5 points or more in the MI Write study (seven to eight months of growth). Tables 5 and 6 show average growth in scores from baseline to follow-up (typically about four to five or seven to eight months apart, depending on the study) for each study's comparison group, which received typical ELA instruction during the measurement period. Both studies included teachers and students in secondary grades from multiple school districts serving high proportions of students who are Black, Latino, and/or experiencing poverty.

Table 5. Average growth on the Self-Efficacy for Writing Scale from baseline to follow-up: MI Write study comparison group

Unadjusted means (standard deviations)	Overall score	Unadjusted means (standard deviations)		
		Ideation	Conventions	Self-regulation
Fall 2021	66.44 (20.50)	61.41 (23.51)	73.75 (20.38)	65.64 (22.89)
Spring 2022	69.82 (20.20)	65.15 (23.80)	77.51 (19.32)	68.51 (22.36)
Difference	3.38	3.74	3.75	2.87

Source: Survey responses from students in MI Write evaluation comparison group who completed student survey at baseline and follow-up (N = 2,487).

Results are based on a 100-point scale ranging from no confidence to complete confidence.

Table 6. Average growth on the Self-Efficacy for Writing Scale from baseline to follow-up: Ecree study comparison group

		Unadjusted means (standard deviations)		
Unadjusted means (standard deviations)	Overall score	Ideation	Conventions	Self-regulation
Fall 2021	n.a.	52.00 (22.04)	70.20 (22.69)	n.a.
Spring 2022	n.a.	52.99 (22.76)	71.26 (22.28)	n.a.
Difference	n.a.	0.99	1.06	n.a.

Source: Survey responses from students in Ecree evaluation comparison group who completed student survey at baseline and follow-up (N = 844).

Results are based on a 100-point scale ranging from no confidence to complete confidence.

n.a = not applicable because study did not measure the component.

S6. The **Implicit Theories of Writing Scale** is a three-item survey measuring students' beliefs about the malleability of their writing ability using a 6-point scale ranging from completely disagree to completely agree. It is an adaptation of Dweck's growth mindset scale (Dweck et al., 1989) with a specific focus on writing, as research has found that adolescents may have content-area specific beliefs.

- Key publication: Limpo and Alves (2017)
- Considerations for reliability: Limpo and Alves (2017) found that the writing scale met the required reliability threshold (internal consistency was 0.81). The study included 196 Portuguese students in grades 7 and 8.
- Relationship to writing performance: Limpo and Alves (2017) found that malleability of writing is positively correlated with writing performance.
- Other considerations for use: This scale might not show change over time within a group of students.

S7. The **Self-Beliefs, Writing-Beliefs, and Attitude Survey** is a 36-item student survey covering six factors: attitude toward writing, beliefs about self as writer, self-concept, self-efficacy, writing beliefs, and overall motivation. This survey uses a 4-point scale ranging from "very different from me" to "a lot like me." It was adapted from the Motivation for Reading Questionnaire (Wigfield & Guthrie 1995) to focus on writing.

- Key publication: Wright et al. (2019)
- Considerations for reliability: Wright et al. (2019) found that the survey met the required reliability threshold (internal consistency was greater than 0.90 for samples and subsamples). Reliability was established on a sample of students in grades 6 through 8 (75 percent of the sample was eligible for free or reduced-priced lunch and 10 percent were English learners).
- Relationship to writing performance: Wright et al. (2019) found that the subscales, subfactors, and the overall scale are positively correlated with measures of writing performance.

- Other considerations for use: In Wright et al.'s (2019) study, teachers read the survey items to students so that reading ability would not affect results. Students completed a paper-based version of the survey in approximately 15 minutes.

S8. The **Writing Disposition Scale** is an 11-item student survey measuring writing confidence, persistence, and passion. This survey uses a 5-point Likert-type scale ranging from strongly agree to strongly disagree.

- Key publication: Piazza and Siebert (2008)
- Considerations for reliability: Piazza and Siebert's (2008) study established reliability with a sample of students in grades 4 and 6 that was majority White (internal consistency was 0.89 for the overall scale), and reliability was not reported for racial or ethnic subgroups.
- Relationship to writing performance: Unknown.
- Other considerations for use: Students completed a paper-based version of the survey in approximately 20 minutes.

Enjoyment of writing

S9. The **Liking Writing Scale** is a four-item student survey measure constructed to provide general information about the extent of students' positive attitudes about writing. Students rate their feelings on a 5-point Likert-type scale ranging from strongly disagree to strongly agree.

- Key publications: Bruning et al. (2013), MI Write study summary (2023), Ecree study summary (2023)
- Considerations for reliability:
  - In Bruning et al. (2013), reliability was established with a sample of 11th- and 12th-grade students that included primarily White students (internal consistency was 0.83).
  - In a study of the automated writing feedback tool MI Write, the four-item scale met the required reliability threshold (internal consistency was 0.84 at baseline and 0.86 at follow-up). The study team removed the midpoint of the 5-point Likert scale, resulting in a 4-point scale (0-3, ranging from strongly disagree to strongly agree). The study took place during the 2021–2022 school year and included 7th- and 8th-grade ELA teachers from three school districts serving high proportions of students who are Black, Latino, and/or experiencing poverty (60 percent of students who participated in the study were Black and/or Latino, and 60 percent were eligible for free or reduced-price lunch).
  - In a study of the automated writing feedback tool Ecree, the four-item scale met the requirement reliability threshold (internal consistency was 0.83 at baseline and 0.83 at follow-up). The study team removed the midpoint of the 5-point Likert scale, resulting in a 4-point scale (1-4, ranging from strongly disagree to strongly agree). The study took place during the 2021–2022 school year and included grade 8–11 ELA teachers from two school districts (22 percent of students who participated in the study were Black and/or Latino).
- Relationship to writing performance: Unknown.

- Other considerations for use: In Bruning et al. (2013), students completed a paper-based version of the survey in approximately 20 minutes. There is little information available about the extent to which student enjoyment of writing normally changes over time. However, in two studies of automated writing feedback tools, MI Write and Ecree, change over time was calculated. At the start of the school year, the mean score in the MI Write study was 1.76 (using a 0-3 Likert scale) and the mean score for the Ecree study was 2.59 (using a 1-4 Likert scale) among students receiving typical ELA instruction.[2] By spring data collection, mean scores decreased by an average by 0.08 points. Tables 7 and 8 show average growth in scores from baseline to follow-up (typically around four to five or seven to eight months apart, depending on the study) for each study's comparison group, which received typical ELA instruction during the measurement period. Both studies included teachers and students in secondary grades from multiple school districts serving high proportions of students who are Black, Latino, and/or experiencing poverty.

Table 7. Average growth on the Liking Writing Scale from baseline to follow-up: MI Write study comparison group

	Unadjusted means (standard deviations)
Fall 2021	1.76 (0.66)
Spring 2022	1.70 (0.68)
Difference	-0.06

Source: Survey responses from students in MI Write evaluation comparison group who completed student survey at baseline and follow-up (N = 2,487).

The results are based on a Likert-type scale from strongly disagree (0) to strongly agree (3).

Table 8. Average growth on the Liking Writing Scale from baseline to follow-up: Ecree study comparison group

	Unadjusted means (standard deviations)
Fall 2021	2.59 (0.62)
Spring 2022	2.50 (0.66)
Difference	-0.09

Source: Survey responses from students in Ecree evaluation comparison group who completed student survey at baseline and follow-up (N = 844).

The results are based on a Likert-type scale from strongly disagree (1) to strongly agree (4).

S10. The **Writing Attitudes Survey** is a 28-question survey for students to complete. It uses a 4-point cartoon scale to rate their feelings about writing in different scenarios.

- Key publication: Kear et al. (2000)

[2] Because the two studies quantified the scales differently (0-3 for MI Write, 1-3 for Ecree), the means of the scale across the MI and Ecree studies cannot be compared to one another. Comparisons can be made, however, between fall and spring scores for the same scale.

- Considerations for reliability: Kear et al. (2000) found that the survey met the required reliability threshold (internal consistency ranged from 0.88 to 0.93 across grade levels). The survey was normed on a nationally representative sample of students in grades 1 through 12, including African American students (16 percent) and Latino students (8 percent).
- Relationship to writing performance: Unknown.
- Other considerations for use: The cartoons might not be appropriate for secondary grade levels, so users might want to adapt the graphics. Individual items from this survey have been used in other research.

Perceived value of writing

S11. The **Writing Achievement Goals Scale** is a 12-item scale for students to complete using a 5-point Likert-type scale ranging from "does not describe me at all" to "describes me perfectly." The survey assesses a three-factor model covering students' intentions or goals in writing as part of their ELA class. The three factors are mastery goals, performance approach goals, and performance avoidance goals.

- Key publication: Soylu et al. (2017)
- Considerations for reliability: Soylu et al. (2017) established reliability for each factor on a sample of 8th-grade, primarily White students (internal consistency was 0.84 for mastery goals, 0.81 for performance approach goals, and 0.75 for performance avoidance goals).
- Relationship to writing performance: Soylu et al. (2017) found that the three subscales are related to writing performance on state tests; however, the relationship for the performance avoidance goals is not statistically significant.
- Other considerations for use: This instrument was designed to understand intentions rather than measure actual outcomes. If examining standardized writing test scores as the outcome, users should consider that students might not see performance on a standardized test as being a writing task within their class.

S12. The **Beliefs about Writing Survey** is a 31-item survey for students to complete using a 5-point Likert scale to rate their beliefs about writing transmission (sticking closely to arguments, information, and quotations provided by authorities), transaction (affective and cognitive engagement), recursive process (an iterative approach to writing), and audience orientation (a focus on the readers and their interests).

- Key publication: Sanders-Reio et al. (2014)
- Considerations for reliability: Sanders-Reio et al. (2014) found that three of the four subscales met reliability standards (internal consistency was 0.72 to 0.98); the transmission subscale did not (internal consistency was 0.65). Reliability was established with a sample of undergraduate college students, so the measure's reliability may perform differently with samples of students from secondary grade levels. The sample was 68 percent Latino and 11 percent Black (reliability was not reported separately by racial or ethnic group).

- Relationship to writing performance: Sanders-Reio et al. (2014) found that audience orientation and recursive process were positively related to writing performance. Transmission was associated with lower writing performance, which may be because students with this belief might use a more mechanical approach to writing.
- Other considerations for use: This survey may not show changes over time, and users should consider pilot testing the measure with their population and whether it is appropriate for their research questions.

Motivation to write

S13. The **Writing Activities and Motivation Scale** is a student survey with two main sections: (1) 30 items related to writing motivation on an 11-point scale ranging from totally disagree to totally agree and (2) 10 items related to frequency of writing activities on a 5-point scale ranging from almost never to almost daily. This measure is also listed in the section on measures of student writing mindsets.

- Key publication: Troia et al. (2013)
- Considerations for reliability: Troia et al. (2013) found that the required reliability threshold was met for the motivational beliefs scale (internal consistency was 0.88) but not for the achievement goal orientation scale. The study included 618 students in grades 4 to 7 and grades 9 and 10 (23 percent of students who participated in the study were Black and/or Latino); about one-third of the participating students were in grades 9 and 10.
- Relationship to writing performance*:* Troia et al. (2013) found that motivational beliefs were positively related to narrative writing performance.
- Other considerations for use: The writing motivational scale includes multiple constructs that can be measured separately by other measures (for example, self-efficacy can be measured with other instruments). Relying on a broad measure of motivation may obscure understanding of key dimensions of motivation. The survey items are not published in the article but are available from the authors.

S14. The **Writing Motivation and Engagement Scale** consists of 44 items and examines writing beliefs across 11 subscales: self-efficacy, valuing, mastery orientation, persistence, planning, task management, anxiety, failure avoidance, uncertain control, self-handicapping, and disengagement. Students respond to the items using a 7-point scale ranging from strongly disagree to strongly agree.

- Key publication: Collie et al. (2016)
- Considerations for reliability: Collie et al. (2016) found that the scale met the required reliability threshold (internal consistency ranged from 0.75 to 0.92 across subscales). The sample included 781 male high school students in Australia with above-average socioeconomic status compared to the average for Australian schools. Individual subscales have been shown to have high reliability, and users could use individual subscales on their own.

- Relationship to writing performance: Collie et al. (2016) found that the adaptive subscales (for example, self-efficacy) were positively related to writing and literacy outcomes, and the maladaptive subscales (for example, self-handicapping) were negatively related to writing and literacy outcomes.
- Other considerations for use: Users might want to select individual subscales to limit the length of the survey. This is a norm-referenced assessment that is available for purchase.

S15. The **Writing Motivation Scale** is a 25-item student survey on seven motivational incentives: curiosity, involvement, social recognition, grades, competition, emotional regulation, and relief from boredom. Students respond on a 4-point scale ranging from very true to not true at all.

- Key publication: Camping et al. (2020)
- Considerations for reliability: Camping et al. (2020) found that the scale met the required reliability threshold (internal consistency exceeded 0.71 across all seven motivational incentives). The sample included 570 students in grades 6 through 8 (51 percent of participating students were Latino and 50 percent were currently or previously classified as English learners).
- Relationship to writing performance: Camping et al. (2020) found that the motivational incentives measured in the survey were not related to writing performance for English learners; there was a small but statistically significant relationship for native English speakers.

Measures of student argumentative writing skills

S16. The **Smarter Balanced Argumentative Performance Task Writing Rubric (Grades 6–11)** uses a 4-point scale to assess three traits of writing: organization/purpose, evidence/elaboration, and conventions for argumentative writing. The first two traits are rated on a 4-point scale from 1 (low) to 4 (high), and Conventions is rated on a 3-point scale from 0 (low) to 2 (high). The overall score takes on a value from 1 to 6 calculated as follows: (organization/purpose score + evidence/elaboration score) / 2 + conventions score.

- Key publications: Smarter Balanced Technical Report, MI Write study summary (2023), Ecree study summary (2023)
- Considerations for reliability: Researchers should assess and establish inter-rater reliability with their own raters. However, essay data from two studies of writing feedback tools, MI Write and Ecree, met the required reliability threshold (inter-rater reliability was greater than 0.70) for 17 of 18 subscore-by-subsample groupings. The studies took place during the 2021–2022 school year and collectively included more than 4,000 students in grades 7 through 11 from five school districts serving high proportions of students who are Black, Latino, and/or experiencing poverty. The studies used the same set of prompts to collect argumentative essays from students; separate prompts were used for each of three grade bands (grades 7 and 8, grades 9 and 10, and grade 11) and at the beginning and end of the study. Baseline and follow-up essays for both studies were scored at one time by the same

set of raters, and study, study condition, and time point for each essay were not shared with raters. Experienced Smarter Balanced raters received training using scoring training materials tailored to each of the six essay prompts, and 10 percent of scoreable essays were scored by two raters to assess inter-rater reliability on each of the three subscores (organization/purpose, evidence/elaboration, and conventions) for each subsample grouping (grade band by time point). The rate of exact matches in scores was 0.74 to 1.00 for all but one subscore-by-subsample group: the conventions subscore for the grade 9 and 10 baseline subsample (exact match rate was 0.69 and adjacent match rate was 0.31).

- Relationship to writing performance: Not applicable since this is a measure of writing performance.
- Other considerations for use: Hand scoring description and protocols as well as anchor paper sets are available. Total estimated time for the performance tasks for grades 6 through 8 is 120 minutes. There are publicly released prompts, and users can select those that address topics relevant to students' lives. The reading level of the source materials associated with the tasks might be too challenging for students who read below grade level. Little information is available about the extent to which students' argumentative writing skills normally change over time. However, in two studies of automated writing feedback tools, MI Write and Ecree, mean overall scores ranged from 3.16 to 3.37 (depending on the study sample) at the start of the school year among students receiving typical ELA instruction. By spring data collection, mean scores increased on average by 0.37 points in one study. In the other study, mean scores decreased on average by 0.07 points, but the study team also observed more essays that could not be scored (for example, because they were too short or off topic) in the spring than in the fall, suggesting that the average decrease in scores may reflect lower effort rather than reductions in skill. Tables 9 and 10 show average growth in scores from baseline to follow-up (typically around four to five or seven to eight months apart, depending on the study) for each study's comparison group, which received typical ELA instruction during the measurement period. Both studies included teachers and students in secondary grades from multiple school districts serving high proportions of students who are Black, Latino, and/or experiencing poverty.

Table 9. Average growth on the Smarter Balance Rubric from baseline to follow-up: MI Write study comparison group

Unadjusted means (standard deviations)	Overall score	Unadjusted means (standard deviations)		
		Purpose/organization	Evidence/elaboration	Conventions
Fall 2021	3.37 (1.23)	1.80 (0.77)	1.79 (0.77)	1.57 (0.63)
Spring 2022	3.74 (1.26)	2.03 (0.84)	2.02 (0.83)	1.72 (0.56)
Difference	0.37	0.23	0.23	0.15

Source: Survey responses from students in MI Write evaluation comparison group who completed student survey at baseline and follow-up (N = 2,487).

The results are based on a 4-point scale.

Table 10. Average growth on the Smarter Balance Rubric from baseline to follow-up: Ecree study comparison group

Unadjusted means (standard deviations)	Overall Score	Unadjusted means (standard deviations)		
		Purpose/organization	Evidence/elaboration	Conventions
Fall 2021	3.16 (1.13)	1.77 (0.68)	1.77 (0.68)	1.38 (0.67)
Spring 2022	3.09 (1.11)	1.58 (0.66)	1.55 (0.64)	1.51 (0.66)
Difference	-0.07	-0.19	-0.22	0.13

Source: Survey responses from students in Ecree evaluation comparison group who completed student survey at baseline and follow-up (N = 844).

The results are based on a 4-point scale.

S17. The College Board's **AP English Language and Composition scoring rubric** is a 6-point rubric used to score the nationwide Advanced Placement (AP) English language free-response question. Points are awarded for thesis, evidence and commentary, and sophistication. The rubrics are designed to be specific to a writing prompt, so they are not useful unless students are following the same writing prompt. The prompts are related to specific pieces of English literature and are not customizable.

- Key publication: Maneckshana et al. (1999) [not publicly available] summarizes measures characteristics for an older version of the rubric, the 1999 AP English Language Scoring Guidelines. The current version of the scoring rubric was revised in [date].
- Considerations for reliability: Researchers should assess and establish inter-rater reliability with their own raters. The required level of inter-rater reliability was previously established with the 1999 version of the AP English Language Scoring Guidelines.
- Relationship to writing performance: No evidence available.
- Other considerations for use: Guidance is provided to teachers, students, and colleges on interpreting the final score but not on specific writing rubric scores. The rubrics are designed to be specific to a writing prompt, so they are not useful unless students are following the same writing prompt. The prompts are related to specific pieces of English literature and are not customizable.

S18. The **PARCC/New Meridian** scoring rubric for writing is a 4-point rubric used for grades 6 through 11.

- Key publication: PARCC Technical Documentation (2017)
- Considerations for reliability: Researchers should assess and establish inter-rater reliability with their own raters; however, the required level of inter-rater reliability has been established with this rubric in other contexts. The 2017 PARCC Technical Documentation found an inter-rater reliability of 0.75 (exact match rate) for the ELA/literacy assessment in grades 3 through 8 and high school during the 2016–2017 academic year.
- Relationship to writing performance: No evidence available.

- Other considerations for use: This rubric is not specific to argumentative writing; it is for research simulation and literary analysis. The writing tasks are standardized text sets and prompts that fall into three categories: narrative writing, research simulation, and literacy analysis.

S19. The **Literacy Design Collaborative (LDC) Student Work Rubric for Argumentation Tasks** for grades 9 through 12 is a 7-point rubric that includes five domains: Controlling Idea, Selection and Citation of Evidence, Development/Explanation of Sources, Organization, and Conventions. This version is aimed at the high school level, with "meets expectations" aligned with state standards for grades 11 and 12. There is also a middle school version aligned with state standards for grade 8. Templates and guidance are available for teachers to write their own writing tasks, and teachers can select tasks from the LDC Core Tools Library.

- Key publication: Wei and Cor (2015)
- Considerations for reliability: Researchers should assess and establish inter-rater reliability with their own raters; however, unpublished research from LDC confirms reliability for the 2016 version of the rubric. Reliability was established for the total score, but not at the trait-level.
- Relationship to writing performance: No evidence available.
- Other considerations for use: The grade 9–12 rubric grading scale (out of 4 points) is aligned to standards for 11th- and 12th-grade students. Therefore, it is expected that students in grades 9 and 10 will have lower scores. Anchor paper sets for understanding each score level are available, as well as training protocols from LDC. Professional development modules and online courses are offered for calibration and training (current rates are $200 per teacher or $3,000 per district).

S20. The **Score Basic Elements** rubric is a 7-point guideline for the essential elements for writing a persuasive essay.

- Key publication: Kiuhara et al. (2012)
- Considerations for reliability: Researchers should assess and establish inter-rater reliability with their own raters; however, the required level of inter-rater reliability has been established with this rubric in other contexts. Kiuhara et al. (2012) found that scores were highly correlated across two raters (correlation was 0.93) in a study of six 10th-grade students.
- Relationship to writing performance: No evidence available.
- Other considerations for use: This rubric is designed to be used for argumentative writing tasks; no writing tasks are provided.

References

Brindle, M., Graham, S., Harris, K. R., & Hebert, M. (2016). Third and fourth grade teacher's classroom practices in writing: A national survey. *Reading and Writing, 29,* 929–954. https://doi.org/10.1007/s11145-015-9604-x

Bruning, R., Dempsey, M. & Kauffman, D., McKim, C. & Zumbrunn, S. (2013). Examining Dimensions of Self-Efficacy for Writing. *Journal of Educational Psychology, 105,* 25–38. **https://psycnet.apa.org/doi/10.1037/a0029692**

Camping, A., Graham, S., Ng, C., Aitken, A., Wilson, J., & Wdowin, J. (2020). Writing motivational incentives of middle school emergent bilingual students. *Reading and Writing, 33,* 2361–2390. **https://doi.org/10.1007/s11145-020-10046-0**

Collie, R., Martin, A. J., & Curwood, J. S. (2016). Multidimensional motivation and engagement for writing: Construct validation with a sample of boys. *Educational Psychology, 36*(4), 771–791. https://doi.org/10.1080/01443410.2015.1093607

Cor, M. K. (2011). *Investigating the reliability of classroom observation protocols: The case of PLATO* [Unpublished manuscript]. Stanford University.

Dweck, C. S., & Henderson, V. L. (1989). Theories of intelligence: Background and measures. Paper presented at the Biennial Meeting of the Society for Research in Child Development, Kansas City, MO.

Gallagher, H. A., Arshan, N., & Woodworth, K. (2017). Impact of the National Writing Project's College-Ready Writers Program in high-need rural districts. *Journal of Research on Educational Effectiveness, 10*(3), 570–595.**https://doi.org/10.1080/19345747.2017.1300361**

Gallagher, H. A., Woodworth, K. R., Wang, H., Bland, J. A., Bosetti, K. R., Cassidy, L. J., Gallagher, L. P., Hafter, A., McCaffrey, T. Murphy, R. F., & Shields, P. M. (2012). *National evaluation of Writing Project school partnerships: Final report.* SRI International. **https://www.sri.com/wp-content/uploads/2021/12/wpd_final_report_execsummary_oct_22.pdf**

Graham, S., Harris, K., MacArthur, C., & Fink, B. (2001). Primary Grade Teachers' Theoretical Orientations Concerning Writing Instruction: Construct Validation and a Nationwide Survey. *Contemporary Educational Psychology. 27.* 147-166. **https://www.researchgate.net/publication/222020615_Primary_Grade_Teachers'_Theoretical_Orientations_Concerning_Writing_Instruction_Construct_Validation_and_a_Nationwide_Survey**

Graham, S., Capizzi, A., Harris, K. R., Hebert, M., & Morphy, P. (2014). Teaching writing to middle school students: A national survey. *Reading and Writing, 27*(6), 1015–1042. **https://doi.org/10.1007/s11145-013-9495-7**

Graham, S., Harris, K. R., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment*. Alliance for Excellence in Education. **https://www.carnegie.org/publications/informing-writing-the-benefits-of-formative-assessment/**

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*, 879–896. **https://psycnet.apa.org/doi/10.1037/a0029185**

Gray, D. L., Hope, E. C., & Matthews, J. S. (2018). Black and belonging at school: A case for interpersonal, instructional, and institutional opportunity structures. *Educational Psychologist, 53*(2), 97–113. **https://doi.org/10.1080/00461520.2017.1421466**

Kear, D., Coffman, G., McKenna, M., & Ambrosio, A. (2000). Measuring attitude toward writing: A new tool for teachers. *The Reading Teacher, 54,* 10–23.

Kiuhara, S., O'Neil, R., Hawken, L., & Graham, S. (2012). The effectiveness of teaching 10th grade students with a disability a strategy for planning/drafting persuasive text. *Exceptional Children, 78*(3), 335–355. **https://doi.org/10.1177/001440291207800305**

Limpo, T., & Alves, R. A. (2017). Relating beliefs in writing skill malleability to writing performance: The mediating role of achievement goals and self-efficacy. *Journal of Writing Research, 9*(2), 97–125. **https://doi.org/10.17239/jowr-2017.09.02.01**

Looney, L. (2003). Understanding teachers' efficacy beliefs: The role of professional community. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 64* (12-A), 4357.

MacArthur, C. A., Philippakos, Z. A., & Ianetta, M. (2015). Self-regulated strategy instruction in college developmental writing. Journal of Educational Psychology, 107(3), 855–867. **http://dx.doi.org/10.1037/edu0000011**

Maneckshana, B., Morgan, R., & Batleman, M. (1999). *Advanced Placement English literature and composition form 3VBP reader reliability study* [Unpublished statistical report]. Educational Testing Service.

Morphy, P., & Graham, S. (2012). Word processing programs and weaker writers/readers: A meta-analysis of research findings. *Reading and Writing, 25,* 641–678 **https://doi.org/10.1007/s11145-010-9292-5**

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. **https://learning.ccsso.org/wp-content/uploads/2022/11/ELA_Standards1.pdf**

Piazza, C. L., & Siebert, C. F. (2008). Development and validation of a writing dispositions scale for elementary and middle school students. *Journal of Educational Research, 101*(5), 275–285. **https://doi.org/10.3200/JOER.101.5.275-286**

Ray, A. B., Graham, S., & Liu, X. (2018). Effects of SRSD college entrance essay exam instruction for high school students with disabilities or at-risk for writing difficulties. Reading and Writing: An Interdisciplinary Journal. **https://doi.org/10.1007/s11145-018-9900-3**.

Sanders-Reio, J., Alexander, P. A., Reio, T. G., & Newman, I. (2014). Do students' beliefs about writing relate to their writing self-efficacy, apprehension, and performance? *Learning and Instruction, 33,* 1–11. http://dx.doi.org/10.1016/j.learninstruc.2014.02.001

Soylu, M. Y., Zeleny, M. G., Zhao, R., Bruning, R. H., Dempsey, M. S., & Kauffman, D. F. (2017). Secondary students' writing achievement goals: Assessing the mediating effects of mastery and performance goals on writing self-efficacy, affect, and writing achievement. *Frontiers in Psychology, 8,* 1406. **https://doi.org/10.3389/fpsyg.2017.01406**

Stanford University. (2013). *The Protocol for Language Arts Teaching Observation (PLATO): Description of the thirteen elements*.

Troia, G. A., Harbaugh, A. G., Shankland, R. K., Wolbers, K. A., & Lawrence, A. M. (2013). Relationships between writing motivation, writing activity, and writing performance: Effects of grade, sex, and ability. *Reading and Writing, 26,* 17–44. **https://doi.org/10.1007/s11145-012-9379-2**

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783–805. https://doi.org/10.1016/S0742-051X(01)00036-1

Wei, R. C., & Cor, K. (2015). *Assessing what matters: Literacy Design Collaborative (LDC) writing tasks as measures of student learning*. Stanford Center for Assessment, Learning, and Equity (SCALE).

Wigfield, A. & Guthrie, J. (1995). Dimensions of Children's Motivations for Reading: An Initial Study. Reading Research Report No. 34. *National Reading Research Center, 34.* **https://eric.ed.gov/?id=ED384010**

Wijekumar, K., Graham, S., Harris, K. R., Lei, P., Barkel, A., Aitken, A., Ray, A., & Houston, J. (2019). The roles of writing knowledge, motivation, strategic behaviors, and skills in predicting elementary students' persuasive writing from source material. *Reading and Writing, 32*(6), 1431–1457. **https://doi.org/10.1007/s11145-018-9836-7**

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *The Journal of Educational Computing Research, 58*(1), 87–125. **https://doi.org/10.1177/0735633119830764**

Wright, K. L., Hodges, T. S., & McTigue, E. M. (2019). A validation program for the Self-Beliefs, Writing-Beliefs, and Attitude Survey: A measure of adolescents' motivation toward writing. *Assessing Writing, 39,* 64–78. https://doi.org/10.1016/j.asw.2018.12.004

Appendix

Appendix A. Criteria and steps for assessing teacher measures

This appendix includes the criteria and the steps for assessing teacher measures. These criteria and steps were used to identify and assess measures in the menu and can be applied to other measures as well. In addition to this set of general criteria, users assessing and selecting measures should consider the context-specific questions listed in the menu in Box 2 in the main document. 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: Determine if measures meet required criteria

Measure has adequate evidence of score reliability (for example, internal consistency, inter-rater reliability)

- For survey measures: Internal consistency for any scale or subscale used as a measure should be 0.70 or higher.
- For observational measures: Inter-rater reliability should be 0.70 or higher. Users should establish inter-rater reliability with their own raters.

Step 2: Determine if measures meet the preferred criteria

Reliability has been established with secondary school teachers

- Prior research should show that the measure can be used reliably in secondary school grade levels.

Measure has clearly defined topics or constructs measured by each subscale

- If a measure has subscales, each one should measure clearly defined and distinct constructs.

Measure is available for use without restrictions on access

- Users should be able to access and use the measure without substantial cost or burden.

Measure is specific to writing

- A measure should be specific to writing beliefs, attitudes, behaviors, or performance. Alternatively, it can be adapted from another discipline to be specific to writing. For example, when asking about beliefs about student ability, measures should focus on teachers' beliefs about students' writing ability rather than more generally about their students' academic ability.

Measure was developed for research purposes

- Measures may be developed for different purposes, such as research, professional evaluation, or for professional development. Using a measure that was developed for research purposes is preferred.

Step 3: Determine if measures meet preferred criteria by specific type of measure

Teacher surveys and logs

Measure includes clear instructions for ease of use

- For teacher logs and observation instruments, clear instructions are needed to ensure that the measure is used consistently across subjects and settings.

Measure gauges both frequency and intensity of focal instructional activities

- A strong measure of instructional practice will assess not only that a practice is used, but also how prominently it is used in instruction. This applies only to measures of teacher instructional activities.

Measure assesses teacher attitudes, beliefs, and practices that are sensitive to change with intervention

- Measures should focus on beliefs, attitudes, and practices that would be expected to change after implementing an intervention. Users should avoid measures focused on aspects of beliefs, attitudes, and practices that are likely to be more stable within teachers over time.

Observational measures

Materials are available for training raters and ensuring that they apply the measure in the same way

- Training materials are available, clear, and actionable.

Measure can be adapted for remote administration

- In light of in-person meeting restrictions due to COVID-19, it may be necessary to administer observational measures remotely.

Users should consider questions in Steps 4 and 5 to determine whether the measure is a good fit for the local context.

Step 4: Consider context-specific questions to assess validity

Does the measure have face validity?

- Is the measure designed to capture a teacher outcome in the intervention's theory of change? A sufficient description of the outcome measure must be provided to determine

that the measure is clearly defined and the content assessed by the measure aligns with its definition. For example, a measure described as an assessment of teacher attitudes toward teaching writing that actually assesses teacher attitudes toward writing in general (not writing instruction) does not have face validity.

- Is the measure aligned with a teacher outcome expected to change at the point of implementation when it is planned to be used? Longer-term outcomes should not be measured until the time at which the intervention would be expected to produce improvements.

Is there evidence that the measure is predictive of expected longer-term student outcomes in the intervention's theory of change?

Step 5: 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?

Appendix B. Criteria and steps for assessing student measures

This set of criteria includes the steps for assessing secondary writing measures. These criteria were used to identify and assess measures in the menu and can be applied to other measures as well. In addition to this set of general criteria, users assessing and selecting measures should 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: Determine if measures meet required criteria

Measure has adequate evidence of score reliability (for example, internal consistency, inter-rater reliability)

- For surveys with self-reported measures of beliefs, attitudes, and behaviors related to the same construct: Internal consistency for any scale or subscale used as a measure should be greater than 0.70.
- For measures that have raters assigning ratings or scores: Inter-rater reliability must be greater than 0.70. Users should establish inter-rater reliability with their own raters.

Measure is specific to writing

- Measure is specific to writing beliefs, attitudes, behaviors, or performance OR is adapted from another discipline to be specific to writing. For example, when asking about beliefs, measures should focus on students' beliefs about their writing ability rather than more generally about academic ability.

Step 2: Determine if measures meet the preferred criteria

Measure is culturally responsive for use with communities in focus

- The measure has demonstrated reliability and validity with students from communities in focus OR has been qualitatively tested (through focus groups, cognitive interviewing, or talk-aloud protocols) with students from communities in focus.
- If the measure was modified to reflect findings from qualitative testing, reliability and validity were re-established on the modified version.

Reliability has been established with secondary school students

Measure has clearly defined topics or constructs measured by each subscale

Measure is available for use by any grantees without restrictions on access

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

Guidance is available to help users consistently interpret measure results

Measure uses student-friendly accessible language

- The reading level of the measure is appropriate for secondary school students based on Microsoft Word or lexile.com.

Step 3: Consider linguistic accessibility

Is the measure linguistically accessible (for English learners who also speak Spanish)?

Directions and test items are presented in Spanish alongside original English version OR one or more tools are provided to support comprehension, such as:

- Audio versions of questions
- Dictionaries
- Pop-up glossaries
- Audio glossaries
- Printed glossaries

Translations and tools include minimal barriers to use (for example, minimal mouse-clicking skills)

Measure has been translated by a culturally competent translator using best practices in translation, including:

- Translator is a native speaker and considers differences in vocabulary, pronunciation, forms of speech, and idiomatic expressions in multiple dialects
- Translator has writing content knowledge (if applicable)
- Translation is performed by independent translators with translation reconciliation to resolve discrepancies

Step 4: Consider context-specific questions to assess validity

Questions to assess validity

- Is the measure designed to capture an outcome in the intervention's theory of change?
- Is the measure aligned with an outcome expected to change at the point of implementation when it is planned to be used? Longer-term outcomes should not be measured until the time at which the intervention would be expected to produce improvements.
- Is there evidence that the measure is predictive of expected longer-term outcomes as defined by the intervention's theory of change?
- Does the design of the measure match the intended use:

- Is the measure designed to be a formative assessment, that is, intended to be used during a unit or course to measure progress and learning? Or is it designed to be a summative assessment, that is, intended to measure what students have learned at a defined end point of a unit or course?
- Does the measure capture growth or proficiency?

Questions to assess context-specific linguistic accessibility and cultural relevance

- Are 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?
- Do participating students speak languages other than Spanish? 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?

Step 5: 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?

Acknowledgments

A panel of twelve researchers in the field of writing instruction and measurement provided guidance on the secondary writing measures summarized in this menu. Panelists included: Linda Friedrich, Karen Harris, Troy Hicks, Nicole Merino, Eugenia Mora-Flores, Detra Price-Dennis, Kay Wijekumar, Steve Graham, Ruth Wei, Eugenia Mora-Flores, Joshua Wilson and Maisha T. Winn. Mathematica staff (Claire Smither Wulsin, Annalee Kelly, and Isabel Callaway) compiled information on the measures for publication. Mathematica staff (Megan Shoji, Marykate Zukiewicz, Lindsay Fox, and Kaleen Healey) also provided feedback on the menu. Jill Miller provided design and production support, and Jennifer Brown provided editorial support. This publication was prepared for 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