Received: 29 November 2023 | Revised: 5 March 2024 | Accepted: 15 March 2024

DOI: 10.1111/ssm.12660

RESEARCH ARTICLE

WILEY | SSMA

Validity evidence of the use of quantitative measures of students in elementary mathematics education

Marsha Ing[1] | **Karl W. Kosko**[2] | **Cindy Jong**[3] | **Jeffrey C. Shih**[4]

[1]School of Education, University of California, Riverside, Riverside, California, USA

[2]School of Teaching, Learning and Curriculum Studies, College of Education, Health and Human Services, Kent State University, Kent, Ohio, USA

[3]STEM Education Department, College of Education, University of Kentucky, Lexington, Kentucky, USA

[4]Department of Teaching and Learning, College of Education, University of Nevada, Las Vegas, Nevada, USA

Correspondence
Marsha Ing, School of Education, University of California, Riverside, 1207 Sproul Hall, Riverside, CA 92521, USA.
Email: marsha.ing@ucr.edu

Funding information
National Science Foundation, Grant/Award Numbers: 1644321, 1644314, 1726543, 1920619, 1920621

Abstract

Quantitative measures in mathematics education have informed policies and practices for over a century. Thus, it is critical that such measures in mathematics education have sufficient validity evidence to improve mathematics experiences for students. This article provides a systematic review of the validity evidence related to measures used in elementary mathematics education. The review includes measures that focus on elementary students as the unit of analyses and attends to validity as defined by current conceptions of measurement. Findings suggest that one in ten measures in mathematics education include rigorous evidence to support intended uses. Recommendations are made to support mathematics education researchers to continue to take steps to improve validity evidence in the design and use of quantitative measures.

KEYWORDS

elementary education, math education, student assessment, validity

Quantitative mathematics measures such as student achievement assessments and surveys have informed large-scale mathematics education policies and practices for well over a century. These data from a wide range of sources including the National Assessment of Educational Progress, SAT and College Board, and military tests of recruits informed the National Commission on Excellence in Education's (1983) report, *A Nation At Risk*. On the basis of this data, the report concluded that "declines in educational performance are in large part the result of disturbing inadequacies in the way educational processes itself is often conducted" (p. 17). Data from survey responses gathered from professionals (such as teachers and principals) were used to make broad recommendations for the field of mathematics education that included organizing the curriculum around problem solving and shifting priorities to focus on programs and classroom activities beyond computational facility (National Council of Teachers of Mathematics, 1980). Similarly, data from quantitative measures have been used to gather evidence and make recommendations on "what works" in terms of education programs, practices and policies (What Works Clearinghouse, 2022). Numerous other examples of how quantitative measures in mathematics education are used to inform policies and practices exist. In cases like the National Council of

Teachers of Mathematics' (1980) *An Agenda for Action*, use of such measures influenced the development of standards that led to significant improvement in children's mathematical learning (NCTM, 2014). Yet, in other cases, the consequences may be dire such as the potential for certain measures barring underrepresented groups to higher education due to shortcomings in the tests themselves (Newman et al., 2022). It is clear that quantitative measures in mathematics education have significant potential to transform the educational landscape.

Given the importance of these measures, over the past two decades, many in the mathematics education community have called for more rigorous validity evidence (Carney et al., 2019, 2022; Hill & Shih, 2009; Lavery et al., 2020). Such measures require the "necessary information to evaluate the overall validity of a study's conclusions" (Flake & Fried, 2020, p. 457). Historically, examination of evidence to support the validity of measures has been scant (Bostic, 2023). Given that the validity of measures used in mathematics education are linked to significant implications in educational policy, practice, equity, and the general well-being of students, a measure that has insufficient validity evidence calls into question a research study or policy document incorporating such a measure. Stated more plainly, if our field is based on data that lacks validity evidence, our very understanding of mathematics education may rest on flawed findings. Given the importance of quantitative measures in mathematics education, the purpose of the present study is to evaluate the prevalence and sources of validity evidence in mathematics education measures of elementary students.

1 | LITERATURE REVIEW

Validity is fundamental to educational research. *The Standards for Education and Psychological Measurement* (AERA, 2014) define validity as, "the degree to which evidence and theory support the interpretations of test scores for proposed use of tests" (p. 11). Despite awareness of this definition of validity, researchers have argued that greater consensus around this definition is needed (Folger et al., 2023; Lederman, 2023; Shepard, 2016). While there is agreement that a test or instrument or measure (we use these terms interchangeably in this article) in and of itself is not valid or invalid (Cronbach, 1988) and that it is more appropriate to examine the validity evidence of inferences made regarding a measure (Haertel, 2013; Kane, 2013) there is less consensus around what validity should include and what it should apply to (Newton & Shaw, 2016). Some researchers focus on the intended use of a measure (see for example Cizek, 2012) while others argue for a practical approach that includes both the intended and unintended consequences that occur when the measure is actually used (see for example Messick, 1995).

Even without consensus on how validity should be operationalized, there is agreement that there should be significant concerns when making inferences with limited or no validity evidence. In other words, intended or not, there are serious implications for research and practice when validity issues are not taken up at all (Pepper, 2020). For example, one may wish to infer that scores from an instrument are indicative of an individual's mathematical knowledge. Yet, there may be little to no evidence of how an individual child mathematically reasoned when engaged with the instrument. Without such evidence, higher scores might not actually equate to greater mathematical knowledge. To support more rigorous research in mathematics education, this article contributes to the conversation from Bostic's recent editorial (2023) in *School Science and Mathematics* and other research on quantitative measures in mathematics education (Battista et al., 2009; Hill & Shih, 2009; Sztajn, 2011) by reviewing validity evidence related to quantitative measures of elementary students used in mathematics education research.

In a historical analysis across 50 years of articles in the *Journal for Research in Mathematics Education*, only 7% of the manuscripts mentioned validity evidence among almost 100 articles (Bostic et al., 2019). Recent efforts in mathematics education have emphasized the critical importance of adhering to criteria, such as the *Standards for Education and Psychological Measurement* (hereafter, the *Standards*), to address validity issues in mathematics education research (Carney et al., 2019, 2022; Lavery et al., 2020). For example, one recommendation is for mathematics education scholars to articulate a clear purpose statement when developing or using an instrument in which the interpretation and use of scores are explicit (Bostic et al., 2019; Krupa, Bostic, & Shih, 2019; Lavery et al., 2020). Another recommendation is the proposed creation of instrument abstracts to assist users in selecting and using appropriate instruments (Carney et al., 2022). Such recommendations are not trivial as insufficient validity evidence for measures used leads to a "lack of necessary information to evaluate the overall validity of a study's conclusions" (Flake & Fried, 2020, p. 457). Rather, research in mathematics education that employs measures lacking in validity may be interpreted as invalid themselves.

When considering measures related to students in mathematics, there is a history of bias in testing that is relevant to current practices (Randall et al., 2022). Thus, it is critical that such measures accurately and fairly

assess students' knowledge and perspectives. For example, discussing mathematics assessment of children with mathematical learning disabilities, Lewis and Fisher (2016) suggested many measures do not distinguish between cognitive and non-cognitive factors that can affect how children respond to questions. Similarly, Walkington et al. (2018) examined data from student performance on the National Assessment of Educational Progress and Trends in International Mathematics and Science Study achievement measures and found that students' mathematics performance was influenced by the variables not related to the construct of mathematics achievement, such as the readability characteristics of the word problems. While there are efforts to consider the evidence related to response processes (see for example, Karabenick et al., 2007; Kosko, 2019; Leighton, 2017), the question remains regarding how common such efforts are in evaluating validity evidence for measures of students in mathematics classrooms, and how much of this evidence is assumed by the researcher.

1.1 | Current approaches to validity arguments

An argument-based approach to validity includes evidence of the inferences made regarding a measure (Cronbach, 1988; Kane, 1992, 2006; Shepard, 1993). Kane (2013) describes the "core idea" of this approach requiring one to "state the proposed interpretation and use explicitly and in some detail, and then to evaluate the plausibility of these proposals" (p. 1). This argument-based approach builds on Toulmin's model (1958) for how arguments are made (1958). This model includes six components: "claim," "data," "warrant," "qualifier," "rebuttal," and "backing." The "claim" is the conclusion whose merits we are seeking to establish. "Data" are the facts we appeal to as foundation for the claim. "Warrants" are the propositions that act as a bridge between the data and the claim. "Qualifiers" are the degree of force which our data confer on our claim in virtue of our warrant. "Rebuttal" occurs when there is acknowledgement of another valid view of the situation and there is a need to expand on these potential limitations and "backing" occurs when there is additional support of the warrant. These six components of Toulmin's model have been used to make and evaluate arguments across different disciplines and scenarios. This model has been applied to a range of scenarios in mathematics education including evaluating launching a problem (González & Eli, 2017), debriefing a lesson study (Groth & Follmer, 2021), teaching word problems (Chazan et al., 2012), and reflecting on practice (Metaxas et al., 2016; Wagner et al., 2014).

Toulmin's model has also been used to evaluate arguments made about claims made about the use of measures to make particular inferences. For example, if the "claim" is that the measure supports inferences about elementary students' understanding of how to represent and solve problems involving addition and subtraction, the "qualifier" is that students probably understand how to represent and solve problems involving addition and subtraction (but we are never 100% confident that we can make this inference). "Data" would include students solving particular items on the measure. The "warrant" is that if students successfully solve these particular items, we can conclude that they have an understanding of how to represent and solve problems involving addition and subtraction. The "rebuttal" is that occasionally students may have the correct answer on the items even when they do not have an understanding of how to solve addition and subtraction problems. For example, students might guess the correct answer by chance or have copied the answer from another student. "Backing" for such "rebuttals" could include having experts evaluating items in terms of how well it measures an understanding of addition and subtraction, conducting cognitive interviews with students to examine whether they are using addition and subtraction to solve these items, and examining relationships between performance on the items with other "data" such as student performance on assignments related to addition and subtraction, teacher assigned grades related to addition and subtraction problems. Thus, this particular model of developing and evaluating arguments is useful to frame validity arguments related to quantitative measures in elementary education.

In mathematics education, there are several examples of this argument-based approach. For example, Bostic and colleagues (Bostic & Sondergeld, 2015); Bostic et al. (2017) developed a measure of problem solving that was intended to make inferences about sixth graders' problem solving abilities. Using different data sources and methods such as a panel of content experts, student cognitive interviews, Rasch modeling and relationships with other variables, Bostic and colleagues gathered evidence to make the claim that the items on the measure engage students in the mathematical behaviors and habits described in the standard for mathematical practice (National Governors Association and Council of Chief State School Officers, 2010). Such claims support the use of scores to gather data on students' problem solving abilities. Bostic and colleagues cautioned that the measure should not be used to make inferences about general mathematics achievement or general problem solving abilities.

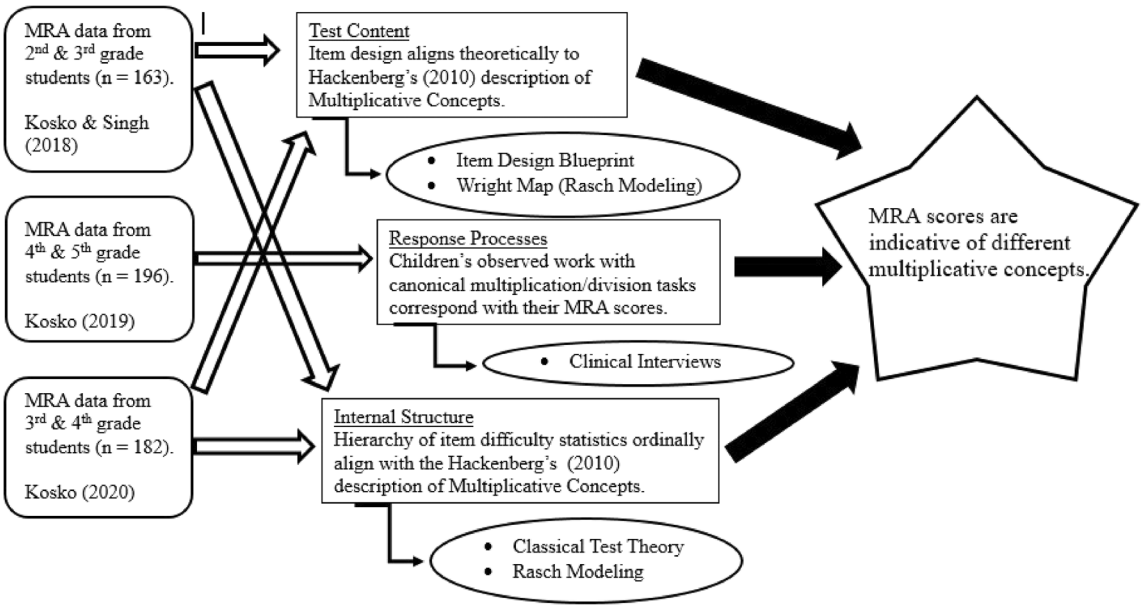

FIGURE 1 Validity argument for one claim for the multiplicative reasoning assessment (MRA).

Similarly, Kosko and Singh (2018, 2019); Kosko (2020) used an argument-based approach to develop a measure of multiplicative reasoning assessment (MRA) in the context of whole numbers for elementary grades. The measure is intended to be used to diagnose students' multiplicative reasoning. Across several publications, Kosko and colleagues described different analyses such as cognitive interviews and Rasch modeling that supported claims such as the MRA can distinguish between groups of students (high, medium, low multiplicative reasoning). For example, one claim regarding the validity of the MRA is that the scores are indicative of different multiplicative concepts (Figure 1). Notably, three different studies provided often overlapping evidence, with different forms of methodological backing (i.e., psychometric statistics and qualitatively examined interviews). Thus, multiple studies can provide various forms of new or confirmatory evidence to support one of several claims for the validity of a measure. Although there are other examples of such an argument-based approach in mathematics education, recent research suggests that this practice may not be widespread (see for example Bostic et al., 2019; Krupa, Bostic, & Shih, 2019; Lavery et al., 2020) which is concerning given the importance of quantitative measures.

2 | CONTEXT OF THE PRESENT STUDY

It is essential that these arguments are established because results from the use of quantitative measures of students in mathematics education are used to inform large-scale policy and practices. For example, the What Works Clearinghouse (2022) reviews research that has used quantitative measures in mathematics and determines which of these findings meet rigorous standards and are "trustworthy" and "meaningful." The Clearinghouse aims to answer the question of "what works in education?" to support users as they identify programs, policies and practices that improve student outcomes. Similarly, the NCTM creates widely read publications such as the *Principles to Action: Ensuring Mathematical Success for All* (NCTM, 2014) that provide guidance to teachers, specialists, coaches, administrators, policymakers and parents. Such publications that shape the field of mathematics education rely on research that has used quantitative measures in mathematics. Funding decisions from federal agencies and foundations and the peer-reviewed publication process also rely on results from quantitative measures in mathematics education. The importance of attending to validity issues related to these measures cannot be overstated given the wide reaching impact.

To elevate the design and implementation of instruments in mathematics education research, it is important to learn more about existing measures and the evidence of validity that are reported. The Validity Evidence for Measurement in Mathematics Education (V-M^{2}Ed) Project aims to create a repository of instruments for mathematics education contexts and their validity evidence. To achieve such a monumental task, sub-teams of researchers focused on a particular population (elementary; secondary; undergraduate/graduate; teacher

education) or content area (e.g., statistics) within mathematics education. Decisions about sub-team membership and the grade bands associated with these sub-teams were made by V-M^{2}Ed leadership. All sub-team members participated in a conference where we worked toward a shared understanding of the two goals of the V-M^{2}Ed Project: (1) to describe, organize, identify, and compile quantitative instruments used in mathematics education contexts and their associated validity evidence; and (2) to develop a framework and conduct syntheses, with the intent to create a publicly available repository of quantitative instruments in mathematics education and their associated validity evidence. Following the conference, sub-teams worked together to achieve these goals. In this article, we report on a systematic search of validity evidence that was collected and analyzed from the sub-team focused on measures for K-6 (elementary) students in mathematics contexts. Our research questions are:

1. What percentage of quantitative instruments focused on elementary students as the unit of analyses that were published in mathematics education research journals include validity arguments?
2. What sources of validity evidence are provided for these instruments?

To address these research questions, our goal is to stimulate conversation about validity evidence in this and other areas of mathematics education, and to encourage others in mathematics education research to attend to measurement issues.

TABLE 1 Outlets included in search.

Outlets included in search
Canadian Journal of Science, Mathematics, and Technology Education
Educational Studies in Mathematics
Elementary School Journal
For the Learning of Mathematics
International Journal for Technology in Mathematics Education
International Journal of Mathematical Education in Science and Technology
International Journal of Science and Mathematics Education
Investigations in Mathematics Learning
Journal for Research in Mathematics Education
Journal of Computers in Mathematics and Science Teaching
Journal of Mathematical Behavior
Journal of Mathematics Teacher Education
Mathematics Teacher Education and Development
Mathematics Teacher Educator
Mathematical Thinking and Learning
Mathematics Education Research Journal
PRIMUS
Research in Mathematics Education
School Science and Mathematics
Teaching Mathematics and its Applications
Technology, Knowledge and Learning
The Mathematics Educator
The Mathematics Enthusiast
ZDM

3 | A SYSTEMATIC SEARCH FOR VALIDITY EVIDENCE

We conducted a systematic search to identify measures used in elementary mathematics education. This involved a review of 1582 articles, with a final sample of 192 measures used in or across research journals specific to mathematics education research, or common for elementary mathematics education research. Following the identification of the measures, we next identified validity evidence and arguments related to the measures. This process is described in depth in the sections that follow.

3.1 | Identifying measures

The first task was to identify elementary (K-6) mathematics measures. We started with an initial set of mathematics education journals that all members of the V-M^{2}Ed Project agreed upon (Williams & Leatham, 2017). We then identified additional outlets that were specific to elementary education, such as *Elementary School Journal*, that might include additional instruments focused on elementary mathematics education. In this first task, we excluded journals that were not were not written in English (Table 1). We also excluded outlets such as proceedings, theses, dissertations, books, and book chapters. Outlets that were not specifically focused on elementary education or mathematics education, such as the *Journal of Educational Psychology* were also excluded. This purposeful decision to focus only on these journals highlights that we were specifically interested in the use of measures in elementary mathematics education research. Our rationale was that if there was an instrument used in a journal not specifically focused on elementary mathematics education (such as the *Journal of Educational Psychology*), it would likely have been used in an elementary mathematics education research journal (so we would have identified that instrument

through our initial search criteria). Instruments that were only used once (such as a measure specifically created for a dissertation) were not likely to include extensive validity evidence and were not likely designed for use by a wider audience. If the measure was used in a dissertation but modified for a wider audience and subsequently used by a broader audience, we would have identified this instrument in our initial search criteria.

Once we identified which journals to search, we then engaged in a process to select search terms to use to identify articles within these journals. This took multiple iterations over several months. Some initial search term language (e.g., "early childhood" AND "instrument") returned results that included articles without quantitative instruments focused on elementary students (e.g., interview protocols for teachers or classroom observation tools). Throughout this process, we were concerned about the replicability and consistency of this process, so the authors tested out a particular set of search terms and reviewed search returns to determine whether the search terms were capturing quantitative measures in elementary mathematics. The sub-team continually discussed any discrepancies in the process. Thus, refinement was necessary and it took a significant amount of attention to detail before a final list of search terms was identified: validity AND sample AND ("children" OR "students") AND ("test" OR "assessment" OR "survey" OR "questionnaire" OR "instrument" OR "measure" OR "scale" OR "protocol" OR "rubric") source: "*Journal for Research in Mathematics Education*."

We searched for instruments used within the date range of 2000–2020. This 20-year range was selected because it included the most recent shift in the *Standards* to focus on interpretation and use statements. The previous version of the *Standards* (1999) also focused on validity in terms of proposed interpretations and uses so we also included years prior to 2014 in our search. This initial search yielded a preliminary list of 1583 articles that included at least one elementary instrument/test in peer-reviewed journal articles and conference proceedings. Over a period of several months, our sub-team individually examined each article to determine whether the source included a measure focused specifically on elementary students. If the instrument included a focus on elementary students, it was flagged for more detailed analysis of validity evidence. Some articles included instruments focused on teachers (such as teacher pedagogical content knowledge) or classrooms (such as classroom observation tools), rather than individual students, and were therefore excluded. For these instruments, the unit of analyses was either teachers or classrooms and not students. This reduced the articles needing further examination to 392. After ongoing discussion where all members of the sub-team coded 20 measures together to come to consensus, the sub-team then individually double coded 10% of the measures to check the consistency of identifying measures through this process. The inter-rater reliability for this initial search process was estimated using Krippendorff's alpha (Krippendorff, 2009) and was found to be high ($\alpha = 0.95$).

Similar to other participants in the V-M^{2}Ed Project, we noted that there were three major categories of instruments based on intended use: broad usage, multi-uses, and single use. Broad-usage instruments were intended for wide use in the field (i.e., the Woodcock-Johnson mathematics assessment; Schrank & Wendling, 2018). Multi-uses instruments were intended for use in future studies by the same researchers who developed the measure or other researchers who were specifically interested in that particular construct, such as the MRA (Kosko, 2019; Kosko & Singh, 2019) or Grade 6 Problem Solving Measure (Bostic et al., 2017; Bostic & Sondergeld, 2015). Single-use instruments were typically designed by researchers and only intended for use in a single study. Given the different intended uses of instruments, the validity evidence varied. For example, although there was little expectation for single-use instruments to include use statements, and offer evidence, it is important for these instruments to include strong validity evidence given their intent to support theory building. In contrast, multiple-uses or broad-usage instruments were expected to include greater validity interpretation, use statements, and evidence.

After the instruments were identified, the sub-team classified instruments by intended use. Almost all (90%) of the broad usage instruments were math achievement-related measures (typically broad definitions of achievement) with 10% being focused on affective factors such as motivation or math anxiety. Multi-uses measures included a higher proportion (29%) that focused on particular concepts such as fractions, probability, and multiplicative reasoning. There was also a greater prevalence of measures focusing on affective factors such as motivation, math anxiety, and peer relatedness (24%). However, a significant portion of multi-uses measures (47%) focused on broader, more generic indicators of mathematics achievement and problem solving. Single-use measures illustrated a similar distribution of focus as multi-use measures: 25% of such measures focused on affective factors and the remaining focused on mathematics concepts and achievement.

3.2 | Categorizing validity evidence

After the instruments were identified, the next step used berrypicking (Bates, 1989) to identify validity evidence and arguments from any article in which the instrument

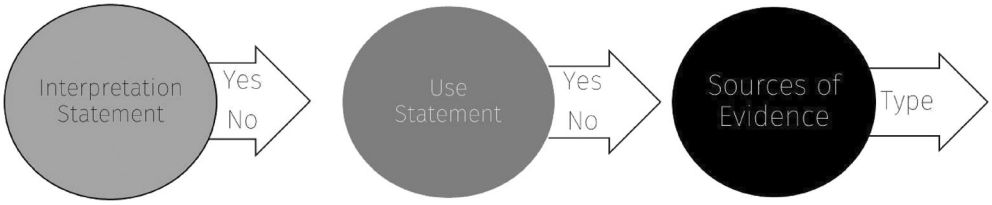

FIGURE 2 Overview of categorizing framework for each instrument.

was used. The sub-team first started with the original article in which the instrument was cited and identified the sources of validity evidence and arguments used in that particular article. The sub-team then searched for other validity evidence related to that instrument from other sources. This search was done by reviewing citations in the article, searching by author or subject in electronic databases. Validity evidence and arguments related to the instruments from these additional articles were then identified and linked with the original instrument used. This process allowed us to search for validity evidence across multiple articles and not rely on the initial article where we identified the instrument. The list of instruments was divided and individually coded by the sub-team. A random sample of 10% of each of the instruments from each author was coded by a second author. Discrepancies were noted and resolved.

3.2.1 | Categorization framework

The validity framework and procedure for categorizing instruments was developed by the V-M^{2}Ed Project leadership team (Bostic et al., 2022; Krupa, Carney, & Bostic, 2019) and shared with conference participants for feedback. After engaging in in-person discussion of the framework, virtual workshops, and whole-group practice applying the framework, each sub-team worked together to apply the framework to their particular instruments (Figure 2). Our sub-team practiced by reading the same articles for an instrument and individually noting whether we could identify an explicit interpretation or use statement and what validity evidence was available across the different articles. Each member of our sub-team independently searched for additional articles and noted references in the articles that might be useful. Once we felt confident to proceed independently, we individually coded the 392 instruments. Approximately 10% of these instruments were double-coded. Discrepancies of the double-coding were flagged and discussed with the whole sub-team. After reaching consensus, the final coding was then recorded in a Google sheet created by the V-M^{2}Ed Project leadership team and checked by two sub-team members.

TABLE 2 Sources of evidence.

Source	Examples
Test content	Alignment with framework or construct Data from experts Literature review
Response process	Cognitive interview Focus groups Written work
Internal structure	Factor analysis Rasch modeling Item response theory
Relations to other variables	Alignment with experts Correlation analysis Triangulation with qualitative data
Consequences of testing	Appropriate cut score Documentation of unintended behavior changes Differential item functioning
Reliability	Generalizability theory Inter-rater reliability Internal consistency or alternatives

Note: Additional information about the sources of evidence provided upon request.

Interpretation statements

The *Standards* refer to validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." (AERA, APA, & NCME, 2014, p. 11). The interpretations of scores for proposed uses are what researchers must evaluate (rather than the test itself). If there is more than one interpretation intended for test scores, each interpretation needs to be evaluated. For example, if one wants to interpret test scores as indicators of teacher effectiveness and student mathematics achievement, each interpretation needs to be evaluated. The interpretation statement includes specification of the construct that the instrument intended to measure. For each instrument, we searched for an explicit statement about how the data from the instrument should be interpreted. If there was an explicit (not implied) interpretation statement, we indicated "yes" and provided a statement from the text. If there was no explicit interpretation statement, then we indicated "no."

Use statements

Similar to interpretation statements, use statements are essential to validity (AERA, APA, & NCME, 2014, p. 11). These statements provide an indication of what the instrument could be used for. Evidence related to the use of the instrument for that particular purpose is then evaluated. For example, the instrument could be used to provide incentives to teachers or used to decide whether students should qualify for a particular intervention. The use statement includes specification of how the instrument is intended to be used. For each instrument, we searched for an explicit statement about how the data from the instrument should be used. If there was an explicit (not implied) use statement, then we indicated "yes" and provided a statement from the text. If there was no explicit use statement, then we indicated "no."

Sources of evidence

The sources of evidence (Table 2) were coded for all the instruments, regardless of whether the instrument included any interpretation or use statement. Many instruments had sources of evidence but were not connected to a particular claim about how the instrument should be used or what interpretations were supported. Many instruments also had multiple evidence for a particular source of evidence. While our coding scheme allowed for capturing multiple evidence for a particular source of evidence (such as if a measure had consequences of testing evidence that included data from experts, differential item functioning, and impacts on clinical/practical implementation), we were primarily concerned with the source of evidence rather than the particular type of evidence within each source. Our search process allowed for gathering information for multiple claims and sources of evidence across several studies. For example, the MRA included evidence for response processes via cognitive interviews in one article (Kosko, 2019), evidence for the association with other variables in another (Kosko & Singh, 2019), and evidence for test content and internal structure in numerous studies. Our coding scheme allowed us to capture these different sources of evidence even though these were found in different articles. In these scenarios, we found that the same authors were primarily responsible for collecting validity evidence and did this across multiple studies.

Throughout this search process, we did not focus on the quality of the evidence provided. We did not evaluate whether the analysis conducted was appropriate or accurate. For example, if the authors indicated that they carried out a factor analysis, then we did not scrutinize whether the type of factor analysis they conducted was appropriate or whether it was accurately conducted or reported. Instead, we identified evidence that the authors used factor analysis and made claims about the structure of the items on the measure in relation to a construct.

4 | RESULTS

A total of 192 distinct measures of elementary students were observed across the sample of 1582 journal articles. As noted previously, several of these measures were observed across multiple articles (21% were broad usage and 28% were multi-study uses), but approximately half (51%) were used in a single study. Despite being the most prevalent form of measure used across the reviewed research, these single-study measures seldom included validity evidence, with the exception of including some indicator of reliability (Table 3). Although 65% included evidence regarding the internal reliability of their measure, approximately a third did not. Rather, single-study measures often included isolated indicators of the technical qualities of their measure, when including validity evidence at all.

Differences in the prevalence of interpretation and use statements across instrument types was notable (Table 3). Although there was validity evidence for 49% of instruments across multiple articles, such that an interpretation and use statement observed in one of many articles for an instrument would be counted toward the instrument as a whole; across all instruments, only 14% included an interpretation statement in at least one article reviewed and 16% included a use statement in at least one article reviewed. These statistics varied by instrument type, which we examined by calculating the Kruskal-Wallis analysis of variance with Dunn's post hoc. We found a statistically significant difference in prevalence of interpretation statements ($H(df = 2) = 25.38$, $p < 0.001$), with single-study use having statistically fewer incorporations than broad ($p < 0.001$) or multi-study use ($p < 0.001$) measures. Similar results were observed regarding prevalence of use statements ($H(df = 2) = 29.74$, $p < 0.001$), with single-study measures having statistically fewer incorporations than broad ($p < 0.001$) or multi-study use ($p < 0.001$) measures, and multi-study measures with fewer incorporations than broad use measures ($p = 0.03$). Despite the differences by instrument type, all observed frequencies are extremely low. Rather, scholars who use quantitative measures of students in mathematics education seldom, if ever, provide an explicit statement for how to interpret the scores from the instrument they use or the intended use of the instrument. This is particularly the case for single-study use instruments, which often intend for such instruments to be used only once. Such instruments are a necessary part of scholarship, but the shortage of these basic forms

TABLE 3 Arguments and evidence across measure types.

	Broad usage	Multi-study uses	Single-study use
Arguments			
Interpretation statement	32%	23%	2%
Use statement	39%	23%	3%
Evidence			
Test content	29%	37%	30%
Response processes	10%	14%	0%
Internal structure	43%	42%	35%
Relations to other variables	42%	40%	17%
Consequence of testing	10%	4%	4%
Reliability	54%	69%	65%

of validity may indicate a lack of transparency in the validity of measures used in mathematics education for elementary students and, perhaps, generally.

Table 3 illustrates the prevalence of sources of validity evidence. Overall, 63% of all measures included evidence toward reliability, with 80% of such instances using Cronbach's alpha as such evidence. Notably, nearly 4 out of 10 instruments did not report reliability evidence, despite our review examining multiple articles for the same measure. Internal structure (41%), test content (33%) and relations to other variables (36%) were the next most common forms of validity evidence, with the latter notably less prevalent for single-study use measures. Each forms of validity included multiple forms of evidence. For example, for instruments including sources of internal structure, 72% used some form of factor analysis, 20% used item response theory (IRT) and 22% used Rasch modeling. Thus, many studies used a combination of these or other statistical analyses. Instruments including relations to other variables most often used correlation analysis (55%) and convergent/divergent association (45%). Test content was most often evidenced by literature reviews (41%), data from experts (38%) and alignments to theoretical frameworks or standards (36%). Both response processes (10%) and consequence of testing (6%) were observed to be rare across the examined measures, with insufficient evidence to highlight a trend in how such validity evidence is collected due to the scarcity of instruments incorporating such forms of validity. More than one in ten measures (13%) presented no validity evidence whatsoever.

While 13% of instruments presented no validity evidence across multiple manuscripts, 25% present a single source of evidence and 37% presented two different sources. These frequencies did not differ by usage type of instrument ($F(df = 2) = 1.50$, $p = 0.23$) and indicated that only a quarter of instruments published in mathematics education literature have provided more than two sources of validity evidence (three sources = 15%, four sources = 6.0%, five sources = 4%). As noted in the prior paragraph, a single source of validity evidence may be supported with more than one type of analysis. For example, several measures included combinations of factor analysis and IRT. Such a combination may provide strong support for internal structure, but may not provide other, necessary, sources of validity evidence.

5 | DISCUSSION

Following our survey of instruments used in mathematics education research to study elementary students, results suggest a severe lack of validity evidence presented for instruments. These results account for multiple publications on the same instrument. Yet, even with such considerations and a scoping review of instruments in 1582 articles, more than 1 in 10 included no validity evidence. Given these results, a relevant question to pose to mathematics education researchers studying elementary students is this: how valid are our interpretations of results if those interpretations were based on instruments lacking validity evidence? As the reader may surmise, this question is predominantly, but not exclusively, theoretical in nature. There is plenty of scholarship expressing the need for improved validity evidence in mathematics education (Carney et al., 2022), though more such work will continue to be needed in research focused on elementary students. It is our hope that the dismal statistics perturb elementary mathematics educators and press the need for action.

Our systematic search of validity evidence in elementary mathematics measures is consistent with recent calls for improving the "standards for the rigorous validation of measures' outcomes and careful attention to the uses of those measures in mathematics education" (Bostic et al., 2019, p. 1; see also Carney et al., 2022; Lavery

et al., 2020). There are numerous instruments in elementary mathematics education that were developed for and used only once for data collection. These single-use instruments often lack cohesive validity arguments. When validity evidence was included, these single-use instruments tended to include pieces of evidence, such as Cronbach's alpha or confirmatory factor analysis, with little or no interpretation of the results. This makes it less likely that others would use the instrument for their own purposes. There is also an open question as to whether the use of these measures support the claims made in publications. The multi-usage or broad-usage instruments identified tended to include more interpretation and use statements relative to single-use instruments. This may be due to the fact that the designers of these instruments intended for their measures to be used by other researchers. Despite the potential use by a broader audience, the percentage of instruments that included more rigorous evidence was relatively low. Another problematic trend is that when measures included more reported statistical analyses, it seldom corresponded with more sources of validity evidence. For example, one measure reported Cronbach's alpha, test–retest reliability, and Kappa statistics, but the combination of such statistics supports only one form of validity (reliability). These results raise questions for the field about expectations for validity evidence for instruments with different intended uses and whether research claims made using these instruments are supported.

6 | CONCLUSION

Our findings focused on elementary students are consistent with broader research in mathematics education advocating for increased validity evidence in the design and use of measures (see for example, Bostic, 2023). The lack of validity evidence indicate that conclusions based on the use of these measures need to be examined further. This is an urgent concern for our field to address because it is foundational to research in mathematics education. Indeed, it is not unlike, or unrelated to, calls for an increase in replication studies (Cai et al., 2018; Eastman, 1975) for the simple fact that it has to do with the methodological rigor of our field. Should mathematics education researchers continue to shirk responsibility toward rigorous scholarship on how children engage in mathematics, we open ourselves up to criticism from others.

Following Carney et al. (2022), we believe a minimum requirement for quantitative measures used in mathematics education research should be the inclusion of an interpretation and use statement. For situations where the measure is intended for single-use, such a statement may be simplified but should be explicit about the study-specific nature of the measure. Such statements can also establish current validity evidence from prior study, thus allowing for more cumulative scholarship on well used measures. They may also serve as justification for replication studies (Cai et al., 2018), as a single source of a particular form of validity evidence often requires confirmation from different samples in the target population. Additionally, these statements may provide additional justification for aspects of an author's study (i.e., relationship to other variables).

6.1 | Recommendations

Readers may wonder what forms of validity evidence to focus on most, in what order, and so forth. The most comprehensive answer is yes. However, it is also the least helpful. Rather, we turn to advice provided by Kane (1992) who suggested focusing on evidence that addressed the most amount of skepticism regarding claims one's measure might include. Kane (1992) noted that a validity claim of a measure "can be questioned because of existing evidence indicated that it may not be true, because of plausible alternative interpretations that deny the assumption, because of specific objections raised by critics, or simply because of a lack of supporting evidence" (p. 530). For example, Kosko and Singh (2018, 2019); Kosko (2020) focused on addressing questions on whether the MRA, which assesses multiplicative reasoning, could do so without the inclusion of explicit multiplication and division symbols. Addressing such questions required certain forms of validity evidence (i.e., response processes & test content) whereas other measures with different critiques may require more attention on other forms of evidence. However, in focusing on forms of validity that answer particular critiques, we caution researchers from considering validity evidence in terms of isolated methods that are not connected to a particular argument.

Some other recommendations for productive movement toward more modern conceptions of validity are to encourage collaborations among researchers with varying areas of expertise. This could include collaborations between those with measurement and mathematics education expertise. Bringing together those who understand and appreciate the value of validity arguments could be done in ways that are mutually beneficial, as it is uncommon for mathematics educators to have a strong measurement background and for a psychometrician to have a background in mathematics education. Funding that incentivizes these multidisciplinary collaborations and

the development and use of quantitative instruments is another way to think about moving the field forward. For example, perhaps those who receive external funding to develop or use instruments could be invited to contribute to the instrument abstract (Carney et al., 2022). Those who are not developing their own instruments but are using instruments cataloged in the instrument abstracts might be invited to provide standardized details such as how the instrument was used and how validity evidence was investigated in their sample. Users are then responsible for contributing to a shared understanding or shared responsibility in the use of quantitative instruments in mathematics education. Another recommendation is to encourage ongoing or long-term efforts to both develop and use measures rather than only develop measures that are convenient for a particular study or sample. To help the field move toward shared understanding and application of modern measurement conceptions, it is also encouraged to consider training and support for users to both create and add on to these instrument abstracts.

Working to ensure that the instruments have appropriate evidence of how they were intended to be used could be accomplished through instrument abstracts proposed by Carney et al. (2022). Consensus around such minimum criteria could lead the field to a more consistent understanding of the quality of instruments and the role that the instruments play in supporting claims, while also recognizing the need for rigor associated with creating and using other types of measures. In addition, work is needed to generate awareness in the field around validity evidence to create a shared understanding around validity. For example, this V-M^{2}Ed Project provided opportunities for mathematics educators to engage in this work together through workshops and collaborative partnerships. A similar type of effort to engage members of mathematics education research journal editorial boards could support the publication of research that is attentive to validity issues. Editors might be encouraged to include editorial board members with measurement expertise and recommend resources in their submission guidelines to support those interested in publishing with their journal to adhere to modern conceptions of validity. Another suggestion is to offer workshops, seminars, or courses that are focused on measurement issues in mathematics education. Such work has the potential to create more nuance around the intended and actual uses of different instruments (Ing et al., 2021). There is a need for the field to make progress toward the appropriate design and use of instruments so that claims made on the basis of those instruments are supported. This is essential to the type of "foundational, theory-building work needed in our field" (Herbst et al., 2022, p. 2). We encourage the mathematics education field to continue to take steps toward improving the validity evidence of instruments used.

ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under grant numbers 1644321, 1644314, 1920619, 1920621. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ORCID

Marsha Ing https://orcid.org/0000-0002-4156-8239

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, *13*, 407–424.

Battista, M., Boerst, T., Confrey, J., Knuth, E., Smith, M. S., Sutton, J., White, D., & Quander, J. R. (2009). Research in mathematics education: Multiple methods for multiple uses. *Journal for Research in Mathematics Education*, *40*(3), 216–240.

Bostic, J., Krupa, E., Folger, T., Bentley, B., & Stokes, D. (2022, October). *Gathering validity evidence to support mathematics education scholarship*. Proceedings of the North American Chapter of the International Group for the Psychology of Mathematics Education.

Bostic, J., Sondergeld, T., Folger, T., & Kruse, L. (2017). PSM7 and PSM8: Validating two problem-solving measures. *Journal of Applied Measurement*, *18*(2), 151–162.

Bostic, J. D. (2023). Engaging hearts and minds in assessment and validation research. *School Science and Mathematics*, *123*, 217–219.

Bostic, J. D., Krupa, E., Carney, M., & Shih, J. (2019). Reflecting on the past and looking ahead at opportunities in quantitative measurement of K-12 students' content knowledge. In J. B. Bostic, E. E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 205–229). Routledge.

Bostic, J., & Sondergeld, T. (2015). Measuring sixth-grade students' problem solving: Validating an instrument addressing the mathematics common core. *School Science and Mathematics Journal*, *115*, 281–291.

Cai, J., Morris, A., Hohensee, C., Hwang, S., Robison, V., & Hiebert, J. (2018). The role of replication studies in educational research. *Journal for Research in Mathematics Education*, *49*(1), 2–8.

Carney, M., Bostic, J., Krupa, E., & Shih, J. (2022). Interpretation and use statements for instruments in mathematics. *Journal for Research in Mathematics Education*, *53*(4), 334–340.

Carney, M., Crawford, A., Siebert, C., Osguthorpe, R. D., & Thiede, K. W. (2019). Comparison of two approaches to interpretive use arguments. *Applied Measurement in Education*, *32*(1), 10–22.

Chazan, D., Sela, H., & Herbst, P. (2012). Is the role of equations in the doing of word problems in school algebra changing? Initial indications from teacher study groups. *Cognition and Instruction*, *30*(1), 1–38.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31–43.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Waine & H. Braun (Eds.), *Test validity* (pp. 3–17). Erlbaum.

Eastman, P. M. (1975). Replication studies: Why so few? *Journal for Research in Mathematics Education*, *6*(2), 67–68.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465.

Folger, T. D., Bostic, J., & Krupa, E. E. (2023). Defining test-score interpretation, use, and claims: Delphi study for the validity argument. *Educational Measurement: Issues and Practice*, *42*(3), 22–38.

González, G., & Eli, J. A. (2017). Prospective and in-service teachers' perspectives about launching a problem. *Journal of Mathematics Teacher Education*, *20*(2), 159–201.

Groth, R. E., & Follmer, D. J. (2021). Challenges and benefits of using Toulmin's argumentation model to assess mathematics lesson study debriefing sessions. *Investigations in Mathematics Learning*, *13*(4), 338–353.

Haertel, E. (2013). Expanding views of interpretation/use arguments. *Measurement: Interdisciplinary Research and Perspectives*, *11*(1–2), 68–70.

Herbst, P., Chazan, D., Crespo, S., Matthews, P. G., & Lichtenstein, E. (2022). How manuscripts can contribute to research on mathematics education: An expansive look at basic research in our field. *Journal for Research in Mathematics Education*, *53*(1), 2–9.

Hill, H. C., & Shih, J. C. (2009). Examining the quality of statistical mathematics education research. *Journal for Research in Mathematics Education*, *40*(3), 241–250.

Ing, M., Chinen, S., Jackson, K., & Smith, T. M. (2021). When should I use this measure to support instructional improvement at scale? The importance of considering both intended and actual use in validity arguments. *Educational Measurement: Issues and Practice*, *40*(1), 92–100.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., De Groot, E., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*(3), 139–151.

Kosko, K. W. (2019). A multiplicative reasoning assessment for fourth and fifth grade students. *Studies in Educational Evaluation*, *60*, 32–42.

Kosko, K. W. (2020). The multiplicative meaning conveyed by visual representations. *Journal of Mathematical Behavior*, *60*, 1–18.

Kosko, K. W., & Singh, R. (2018). Elementary children's multiplicative reasoning: Initial validation of a written assessment. *The Mathematics Educator*, *27*(1), 3–22.

Kosko, K. W., & Singh, R. (2019). Children's coordination of linguistic and numeric units in mathematical argumentative writing. *International Electronic Journal of Mathematics Education*, *14*(2), 275–291.

Krippendorff, K. (2009). *Testing the reliability of content analysis data. The content analysis reader*. Sage Publications.

Krupa, E., Bostic, J. D., & Shih, J. (2019). Validation in mathematics education: An introduction to quantitative measures of mathematical knowledge: Researching instruments and perspectives. In J. B. Bostic, E. E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 1–13). Routledge.

Krupa, E. E., Carney, M., & Bostic, J. (2019). Argument-based validation in practice: Examples from mathematics education. *Applied Measurement in Education*, *32*(1), 1–9.

Lavery, M. R., Bostic, J., Kruse, L., Krupa, E., & Carney, M. (2020). Argumentation surrounding argument-based validation: A systematic review of validation methodology in peer-reviewed articles. *Educational Measurement: Issues and Practice*, *39*(4), 116–130.

Lederman, J. (2023). Validity and racial justice in educational assessment. *Applied Measurement in Education*, *36*(3), 242–254.

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. Oxford University Press.

Lewis, K. E., & Fisher, M. B. (2016). Taking stock of 40 years of research on mathematical learning disability: Methodological issues and future directions. *Journal for Research in Mathematics Education*, *47*(4), 338–371.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Metaxas, N., Potari, D., & Zachariades, T. (2016). Analysis of a teacher's pedagogical arguments using Toulmin's model and argumentation schemes. *Educational Studies in Mathematics*, *93*(3), 383–397.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. *The Elementary School Journal*, *84*(2), 112–130.

National Council of Teachers of Mathematics. (1980). An agenda for action: Recommendations for school mathematics of the 1980s.

National Council of Teachers of Mathematics. (2014). Principles to actions: Ensuring mathematical success for all.

Newman, D. A., Tang, C., Song, Q. C., & Wee, S. (2022). Dropping the GRE, keeping the GRE, or GRE-optional admissions? Considering tradeoffs and fairness. *International Journal of Testing*, *22*(1), 43–71.

Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 178–197.

Pepper, D. (2020). When assessment validation neglects any strand of validity evidence: An instructive example from PISA. *Educational Measurement: Issues and Practice*, *39*(4), 8–20.

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, *27*(2), 170–178.

Schrank, F. A., & Wendling, B. J. (2018). The woodcock–Johnson IV: Tests of cognitive abilities, tests of oral language, tests of achievement. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 383–451). The Guilford Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*(1), 405–450.

Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 268–280.

Sztajn, P. (2011). Standards for reporting mathematics professional development in research studies. *Journal for Research in Mathematics Education*, *42*(3), 220–236.

Wagner, P. A., Smith, R. C., Conner, A., Singletary, L. M., & Francisco, R. T. (2014). Using Toulmin's model to develop prospective secondary mathematics teachers' conceptions of collective argumentation. *Mathematics Teacher Educator*, *3*(1), 8–26.

Walkington, C., Clinton, V., & Shivraj, P. (2018). How readability factors are differentially associated with performance for students of different backgrounds when solving mathematics word problems. *American Educational Research Journal*, *55*(2), 362–414.

What Works Clearinghouse. (2022). *What works clearinghouse procedures and standards handbook, version 5*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Williams, S. R., & Leatham, K. (2017). Journal quality in mathematics education. *Journal for Research in Mathematics Education*, *48*(4), 369–396.

How to cite this article: Ing, M., Kosko, K. W., Jong, C., & Shih, J. C. (2024). Validity evidence of the use of quantitative measures of students in elementary mathematics education. *School Science and Mathematics*, 1–13. https://doi.org/10.1111/ssm.12660