K-12 Student and Teacher Math Measures: What's out there? What do we need?

Jonathan D. Bostic, Erin Krupa, & Cindy Jong
Bowling Green State Univ., North Carolina State Univ., Univ. of Kentucky
bosticj@bgsu.edu , eekrupa@ncsu.edu , cindy.jong@uky.edu

Introduction

Recent calls to action focus on using educational tools that promote mathematics learning through evidence-based and equity-forward practices (NCTM, 2018). These practices may be derived from scholarship that examines factors related to mathematics teaching and learning using quantitative measures. A purpose of this presentation is to highlight areas of strength and opportunity related to the use of quantitative measures in scholarship examining K-12 mathematics settings. One outcome from this research-in progress is that scholars may become more aware of quantitative assessments for use in their research. A second outcome from this research is to foster conversations among colleagues around collaborative scholarship as well as areas for growth within mathematics education assessment. As a result, scholars may be better equipped to engage in quantitative research within mathematics contexts. Recognizing what is available and relevant to a desired area of study has potential to address contexts connected to topics described in *Catalyzing Change* (NCTM, 2018, 2020, 2020). That is, scholars cannot quantitatively measure constructs described in *Catalyzing Change* until it is known what measures are available and what they assess. This research-in progress aims to engage researchers in ongoing research and promote discussions across attendees.

Relevant Literature

Student Measures

Mathematics education scholarship has a more than 50-year history of using quantitative measures (Bostic et al., 2019). Early research on students' outcomes focused on problem solving (e.g., Post & Brennan, 1976) and content (e.g., Sellke et al., 1991; Shumway et al., 1981). Broadly speaking, these measures were general measures intended to assess a construct broadly (e.g., problem solving, mathematics knowledge of middle school students). As time passed, scholars started to address more specific topics. Examples include the Research-based Early Math Assessment (REMA; Clements et al., 2008) and the Probabilistic Reasoning Questionnaire (PRQ; Primi et al., 2014). The PRQ was designed to provide a measure for students' basic understanding of probabilistic reasoning skills to identify difficulties to help them become more successful in introductory statistics. Also, assessments like the 3D Geometry Thinking Test (Pittalis and Christou, 2010), which identifies middle school students' reasoning with spatial ability, highlight how assessments have become more focused on specific mathematics concepts over the years. As more and more instruments are being developed for very specific purposes, it is important to understand what instruments have been developed, for what purposes have they been developed, and the intended population for that instrument.

Teacher Education

In a similar fashion, assessments related to teacher education contexts have undergone a journey influenced by local and national policies (Bostic & Sondergeld, 2015) as well as exploring the degree to which levels of quality and/or implementation of an intervention are present (Bostic et al., 2021). Quantitatively focused scholarship examining teacher- and instructional-factors has not necessarily been linear (Bostic, 2017) yet it does reflect the influence of research over time. For example, amounts of time that teachers spend on a measurable attribute (e.g., Brophy, 1986), presence of a intervention (e.g., Slavin et al., 2009), and changes in practice as seen in self-report (e.g., Swafford et al., 1997) represent one side of this trajectory within teacher education scholarship. On the other side, informal observation techniques such as anecdotal data (e.g., Farmer et al., 2003), formal observations with a holistic score (e.g., Schoen et al., 2003), and formal observations with multiple specific scores drawn from numerous indicators (e.g., Hill et al., 2012) represent another approach to classroom instruction and teacher-level attributes. In addition to these approaches, there are numerous constructs within teacher education that move beyond observed classroom instruction (e.g., mathematics self-efficacy and mathematics anxiety) that continue to be examined.

With the plethora of instruments used in mathematics teacher education and in K-12 mathematics contexts with students, it becomes critical to know more about the instruments that have been used and areas where mathematics education, as a field of scholars, has opportunities for growth. This proposal seeks to unpack those areas for attendees and promote conversations across colleagues with the intent of igniting partnerships.

Context

Over the last four years, a group of 39 scholars collaborated with the intention to explore literature between 2000-2020 with the goal of locating quantitative instruments used within mathematics education contexts. Scholars included mathematics educators, psychometricians, special educators, and policy experts who had previously conducted quantitative mathematics or statistics education assessment work. All had experience using, creating, and/or validating measures for use within mathematics or statistics education contexts. Scholars formed synthesis teams, which included four to seven individuals. This research report describes a subset of work from the larger group, with findings from the elementary (K-6), secondary (7-12), and teacher education assessment teams.

Data Collection and Analysis

Data collection and analyzes processes are summarized here; more details are provided in Bostic et al. (2022). The PRISMA statement guided the literature search (Rethlefsen et al., 2021). Scholars agreed to use the top 24 mathematics education journals (Williams & Latham, 2017) as the basis for their literature search of quantitative instruments. Articles that included quantitative instruments were culled to create a list of instruments. Instrument names, construct(s) measured by the instrument, and keywords related to the instrument were drawn from them. A format for data collection was also informed by past reviews of literature conducted by synthesis leaders (e.g., Bostic et al., 2021; Bostic & Sondergeld, 2015). Synthesis teams agreed on a format for their literature search such that common search terms describing a population (e.g., teach* and learn*) as well as instrument language (e.g., measur*, test*, and assess*) were

identical. Truncated language, wild card (e.g., *), and logic terms were used to generate a sample space for analysis. Google Scholar, EBSCO, ProQuest, JSTOR databases were used to gather articles for analysis. The largest possible set of articles to analyze was desired to best understand the quantitative instruments used within student and teacher scholarship. As an example, 300 instruments were garnered from more than 3,000 articles describing teacher education scholarship.

Synthesis teams were provided with training from the project leaders prior to data analysis. Following that training, each team conducted their own work to develop interrater agreement prior to coding independently. Each team met the minimum rater agreement, $\kappa > .8$ (Landis & Koch, 1977). By meeting or exceeding this threshold, raters' abilities to describe assessment might be viewed as near perfect agreement (Landis & Koch, 1977).

Groups gathered assessments and categorized them using a shared framework. The framework asked users to systematically input information about the journal article including where it is described (e.g., citation and abstract), the instrument's name, construct, and population of interest. Each synthesis team then reviewed the information for possible issues (e.g., duplicates and misinformation) as well as affirming that measures aligned with the keywords, tags, and other descriptors associated with the measure. This work was collectively, qualitatively analyzed by three teams (e.g., elementary, secondary, and teacher education instruments). After the teams agreed that constructs were effectively described, then each set of terms was analyzed for frequencies within the sample. Teams analyzed constructs for frequencies with a keen goal of understanding areas of which there were greater numbers of assessments measuring similar constructs. In total, synthesis teams found 192 measures related to elementary mathematics contexts, 380 measures related to secondary mathematics contexts, and 284 measures related to teacher education contexts.

Findings

Frequencies for the top five terms describing a construct are shared in relation to each context as well as the percentage observed within the relevant sample space. Instruments associated with these terms will be shared in the presentation. Related to instruments used within elementary mathematics contexts, the top five terms were achievement (n=34; 17.7%), number sense (n=12; 6.3%), geometry and measurement (tie: n=9; 4.7%), fraction (n=8; 4.2%), and attitude, problem, and self-efficacy (tie: n=6; 3.1%). Some terms were observed two or fewer times. Examples included language, pattern, and integer. The majority of instruments measured general achievement. There were very few measures of students' affective characteristics (i.e., attitude and self-efficacy). About 18.3% of the top five terms were instruments devoted to measuring specific mathematics content at the elementary mathematics grade level.

Related to instruments used within secondary mathematics contexts, the top five terms were achievement and knowledge (tie: n=66; 17.3%), algebra (n=46; 12.0%), beliefs, motivation and attitude (tie: n=44; 11.5%), geometry (n=33; 8.6%), and number (n=11; 2.9%). Terms such as behavior, misconception, and quadratic appeared two or fewer times in the sample. With the large-scale state and national assessments, the majority of instruments (34.7%) at the secondary level were general achievement or knowledge measures. Twenty percent of the measures were related to specific content (algebra n=46, geometry, n=33) within mathematical domains. In addition, 23% of the

secondary instruments were related to measuring students' affective characteristics or perceptions (i.e., beliefs, motivation, attitude).

For teacher education instruments, the top five terms were beliefs (n=35; 12%), attitudes (n=19; 6.7%), instructional quality (n=13; 4.6%), science (n=8; 2.8%), and nature of mathematics (n=7; 2.5%). Several words appeared two or fewer times in the sample space including but not limited to: standards, disposition, and orientation. Some instruments claimed to measure topics that spanned mathematics and science teaching practices; hence, science arose frequently in the sample space. Similar to K-12 student contexts, approximately 19% of instruments were designed to measure teachers' affective characteristics (e.g., beliefs or attitudes).

Discussion

We perceive the findings as critical to scholarly discussions for two reasons. First, we noticed that broad notions of achievement were common within K-12 student contexts. This informs current scholars that broad measures of knowledge/achievement are available for use. Related to teacher education, it was evident that there are numerous measures related to beliefs and attitudes, which include self-efficacy. This is fortunate as scholars have options for measures. On the other hand, it may be problematic because such a high number of options may make it difficult to select a measure that is most germane to their contexts for teacher beliefs and attitudes, especially if terms are used synonymously but operationalized in distinct ways. A second observation was that these findings indicate trends in quantitative assessment. While some topics have high frequencies (e.g., attitudes) across K-12 student and teacher contexts, our team observed low frequencies for some topics that seem relevant. For example, we observed low frequencies related to language (elementary), behavior (secondary), and standards (teachers). We hold these findings tentative because they describe frequencies related to the construct. It is plausible that a measure might assess teachers' implementation of teaching standards, as one example, but that language was not used by instrument users. To that end, our team believes it is important to further study the measures and understand (a) how instrument developers believe their tools should be used, (b) the ways in which those instruments are used, and (c) how those instruments are described in the literature. Authors (2022) suggest that instrument developers use an instrument use summary to convey relevant information about the purpose of a measure so that users are better equipped to make effective decisions for their scholarly and teaching needs.

Attendee engagement

We propose to structure our session around a central focus with two goals. That focus is to present our study as a means to stimulate conversations among attendees. Our first goal is to provide them with information about what is currently available related to instruments published and used within the last 20 years. We plan to use approximately 18 minutes to describe the results across the three contexts. In addition to these results described in this proposal, we will share the names of assessments for attendees to use. Our second goal is to provide space for attendees to organize into three groups: elementary, secondary, and teacher education, and spend 12 minutes for discussions. Each group will be led by one speaker from each team who will facilitate discussions. Initial questions to ignite conversations are: What are you/your teams examining using quantitative measures? What constructs or measures do you want to

learn more about? What are challenges that have come up when designing or selecting a quantitative measure? How can use of quantitative tools in your scholarship foster evidence-based, equity-forward outcomes? Depending on the size of attendees, we may sort attendees into sub-groups within one of the three groups so that they may discuss shared interests with each other and form groups that want to create collaborations and partnerships.

Acknowledgement

Ideas in this work stem from multiple grant-funded research studies supported by the National Science Foundation (NSF# 1920621; #1920619). Any opinions, findings, conclusions, or recommendations expressed by the authors do not necessarily reflect the views of the National Science Foundation.

References

Bostic, J., Krupa, E., Folger, T., Bentley, B., & Stokes, D. (2022, October). *Gathering validity evidence to support mathematics education scholarship*. In A. Lischka, E. Dyer, E., R. Jones, J. Lovett, J. Strayer, & S. Drown (Eds.), Proceedings of the forty-fourth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (pp. 100-104). Nashville, TN.

Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education, 24, 5-31, https://*doi.org/10.1007/s10857-019-09445-0.

Bostic, J., Krupa, E., & Shih, J. (2019). Introduction: Aims and scope for Assessment in mathematics education contexts: Theoretical frameworks and new directions. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 1-11). New York, NY: Routledge.

Bostic, J. (2017). Moving forward: Instruments and opportunities for aligning current practices with testing standards. *Investigations in Mathematics Learning, 9*(3), 109-110. https://doi.org/10.1080/19477503.2017.1325662

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*. https://doi.org/10.1111/ssm.12130

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, *41*, 1069.

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: the Research-Based Early Maths Assessment. *Educational Psychology*, *28*(4), 457-482.

Farmer, J. D., Gerretston, H., & Lassak, M. (2003). What teachers take from professional development: Cases and implications. *Journal of Mathematics Teacher Education, 6*, 331-360.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*, 56-64.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

National Council of Teachers of Mathematics. (2018). *Catalyzing change in high school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2020). *Catalyzing change in middle school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2020). *Catalyzing change in early childhood and elementary school mathematics.* Reston, VA: Author.

Post, T. R., & Brennan, M. L. (1976). An experimental study of the effectiveness of a formal versus an informal presentation of a general heuristic process on problem solving in tenth-grade geometry. *Journal for Research in Mathematics Education*, *7*(1), 59-64.

Pittalis, M., & Christou, C. (2010). Types of reasoning in 3D geometry thinking and their relation with spatial ability. *Educational Studies in mathematics, 75*(2), 191-212.

Primi, C., Morsanyi, K., & Chiesi, F. (2014). *Measuring the basics of probabilistic reasoning: The IRT-based construction of the probabilistic reasoning questionnaire.* Proceedings of the 9th International Conference on Teaching Statistics (ICOTS), Flagstaff, Arizona, USA.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic reviews*, *10*(1), 1-19.

Schoen, H. L., Cebulla, K. J., Finn, K. F. & Fi, C. (2003). Teacher variables that relate to student achievement when using a standards-based curriculum. *Journal for Research in Mathematics Education, 34*, 228–259.

Sellke, D. H., Behr, M. J., & Voelker, A. M. (1991). Using data tables to represent and solve multiplicative story problems. *Journal for research in mathematics education*, *22*(1), 30-38.

Shumway, R., Wheatley, G., Coburn, T., White, A., Revs, R., & Schoen, H. (1981). Initial effect of calculators in elementary school mathematics. *Journal for Research in Mathematics Education, 12(*2), 119-141.

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best evidence synthesis. *Review of Educational Research, 79*, 839-911.

Swafford, J. O., Jones, G. A., & Thornton, C. A. (1997). Increased knowledge in geometry and instructional practice. *Journal for Research in Mathematics Education, 28*, 467-483.

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