Paper to be presented at AERA 2022

Measures of Mathematics Teachers' Behavior and Affect: An Examination of the Assessment Landscape

Melissa A. Gallagher, Emanuelle Bardelli, Tim Folger, Adrian Neely, Jonathan Bostic, Temple Walkowiak, Annie Wilhelm, Jeremy Zelkowski*
*All authors contributed equally

Purpose

Although the paradigm wars between quantitative and qualitative research methods and the associated epistemologies may have settled down in recent years within the mathematics education research community, the high value placed on quantitative methods and randomized control trials remain as the gold standard at the policy-making level (USDOE, 2008). Although diverse methods are valued in the mathematics education community, if mathematics educators hope to influence policy to cultivate more equitable education systems, then we must engage in rigorous quantitative research. However, quantitative research is limited in what it can measure by the quantitative tools that exist. In mathematics education, it seems as though the development of quantitative tools and studying their associated validity and reliability evidence has lagged behind the important constructs that rich qualitative research has uncovered. The purpose of this study is to describe quantitative instruments related to mathematics teacher behavior and affect in order to better understand what currently exists in the field, what validity and reliability evidence has been published for such instruments, and what constructs each measure.

1. How many and what types of instruments of mathematics teacher behavior and affect exist?
2. What types of validity and reliability evidence are published for these instruments?
3. What constructs do these instruments measure?
4. To what extent have issues of equity been the focus of the instruments found?

Theoretical Framework

The *Standards* ([AERA et al.], 2014) consider assessment to be a broader term encompassing "any systematic method of obtaining information, used to draw inferences about characteristics of people, objects, or programs; a systematic process to measure or evaluate the characteristics or performance of individuals, programs, or other entities, for purposes of drawing inferences" (p. 216.). On the other hand, a test is "an evaluative device or procedure in which a systematic sample of a test taker's behavior in a specified domain is obtained and scored using a standardized process" (AERA et al., 2014, p. 224). This submission uses the terms assessment and instrument interchangeably and focuses on them, rather than tests which tend to address knowledge in various capacities (e.g., content knowledge or pedagogical content knowledge). We draw upon a shared definition of validity: "the degree to which evidence and theory support interpretations of test scores for proposed uses of tests… The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself" (AERA et al., 2014, p.11). Validity provides a link from the assessment to the results and interpretations from it (Kane, 2013). The degree to which those results and interpretations are believed and valued is connected to the quality and quantity of validity evidence associated with an instrument.

The *Standards* (2014; 1999) provide clear guidelines regarding measurement validity and reliability. They describe five sources of validity evidence that should be addressed to some

degree within a validation argument. Those sources are (a) test content, (b) response processes, (c) relationship to other variables, (d) internal structure, and (e) consequences from testing and bias. Characterizations of those five sources and reliability are provided in Table 1. At a minimum, sufficient validity evidence for the five sources should be collected and shared with transparency in addition to reliability evidence (AERA et al., 2014; 1999). Unfortunately, "evidence of instrument validity and reliability is woefully lacking" (Ziebarth et al., 2014, p. 115) in the literature. There is often an over-reliance on reliability, such as Cronbach's Alpha, as an indicator of validity (Carney et al., in press). Reliability plays an important role; however, "the general rule [is] that reliability is a necessary but not sufficient condition for validity" (Kane, 2013, p. 29).

Furthermore, it is important not to ignore the social implications that accompany the use of a test. Messick (1995) is careful to point out that social consequences of use are important to evaluate for validity. While this paper does not analyze the test use claims, examining the relations between social consequences of use and score interpretation provides an opportunity to discuss what is considered valid and for whom. Thus, identifying sources of validity and reliability evidence of contemporary quantitative instruments is an initial step for interrogating how equity emerges in the research and measurement landscape.

Recently, questions have been raised about the availability of quantitative instruments for use within mathematics teaching contexts including behavior and affective constructs (Boston et al., 2015; Bostic et al., 2021). Few relevant syntheses are available that communicate what tools are available to study such constructs. Without such syntheses, researchers must conduct literature searches to locate a viable quantitative instrument or develop one on their own. Thus, this study seeks to explore the instruments that have been developed to study preservice and inservice teachers' behavior and affect. Responding to this question may provide scholars with information about previously developed relevant tools for mathematics contexts. Prior syntheses of teacher education tools available for use within mathematics teaching contexts have highlighted a lack of attention to validity (Bostic et al., 2021, Bostic et al., 2019). As an example, one synthesis of classroom observation instruments used in peer-reviewed published mathematics education journals between 2000-2017 indicated that only 44% of the 107 classroom observation instruments describe any validity evidence. Only 26% of that same sample had evidence related to two or more validity sources (Bostic et al., 2021). It was common for authors to provide reliability evidence without any contextualization of validity evidence. Two important findings from this synthesis were (a) what classroom observation tools were available and (b) what validity evidence for their scores and use were available. The present study expands this finding by including more teacher education instruments, specifically those that are intended to collect information about preservice or inservice teachers' behavior or affect.

Methods and Data Sources

We are part of a larger research team engaged in documenting and cataloging measures/ instruments/assessments used in mathematics education research. The focus of our team is systematically capturing measures of teachers' affect and behavior in research published from 2000-2020. As primarily quantitative researchers who focus on measure development, the replicability, reliability, and validity of the steps we engaged in was of paramount importance. We highlight our parallel use of the steps described by Thunder and Berry (2016) to enhance the

replicability of our procedures as well as to ensure that we were reliably identifying measures within the articles we reviewed.

Step 1: Determine a Research Question

The research questions were established based on the purpose noted previously.

Step 2: Determine Search Terms

We began by determining our search terms and creating a Boolean search string. We decided to use: teach* OR instruct* OR class* to pull articles related to teachers and teaching. We added the AND function and the terms observ* OR instrum* OR survey OR log OR assess* OR protocol* OR rubric* OR tool* in order to find instruments. The last element of our Boolean search string was the name of the journal article. We searched through each of the 24 journals identified by Nivens and Otten (2017) as being quality mathematics education research journals.

Step 3: Search the Databases

We used the Boolean search string described above to search each of the 24 journals. We exported our findings to an Excel spreadsheet, including the authors' names, the journal name, the article title, and the abstract. This search returned 2286 articles.

Step 4: Select Initial Relevant Studies

We divided up the 2286 articles into sets of about 200 and coded the title and corresponding abstracts to see if they met our inclusion criteria: empirical, a study of teachers' affect or behavior, and any possibility of quantitative data collection or analysis. Each author then coded their set of 200 plus an additional set of about 100 abstracts, which were double-coded. After this round of analysis we were left with 711 articles that met our inclusion criteria.

In the previous round of coding we had only examined the title and abstracts to determine if the article met inclusion criteria, thus the next step was to review the full-text of those 711 articles. We divided these articles into sets of 20. Each author then coded one set of 20 and an additional 8 on which they were the secondary coder. We coded based on the following inclusion criteria: empirical, a study of teachers' affect or behavior, and whether an instrument was used quantitatively. At this step each coder also noted the construct, type of instrument, and other relevant instrument details. After each author had coded one set of 20+8, we calculated kappa reliabilities and discussed any discrepancies in coding across team members. We engaged in four rounds of coding in this way and generated a final list of 287 articles, which was further reduced to 271 distinct instruments.

Steps 5-7: Assess Quality, Synthesize, and Publish

Given the primary purpose of this study was to catalog measures of mathematics teacher behavior and affect, we chose not to assess the quality of the instruments used, but rather to synthesize the validity and reliability evidence related to each for other researchers to evaluate and consider in using such instruments in the future.

Step 6: Synthesize

We worked in teams of two to catalog the 271 instruments for: evidence of validity & reliability using the *Standards* framework (AERA et al., 2014). Each researcher individually searched not just initially found articles, but any articles cited in the found article, as well as Google Scholar for articles that cited the found article and were written by the same authors. We searched each article for validity and reliability evidence related to the instrument. Pairs of researchers met to reconcile their findings repeatedly throughout this process. To date we have cataloged approximately 135 instruments.

Step 7: Publish

We publish our findings to date below.

Results

From the 2286 articles returned from the initial search, we found 271 different measures of teacher behavior or affect published in the 24 mathematics education journals we reviewed. Of these, the most common were surveys (n = 83) and observation protocols (n = 22) (see Table 2). Least commonly found were diagnostic (n = 1) and formative assessments (n = 2).

We found that authors most frequently reported evidence related to test content (n = 40) and internal structure (n = 33). Most authors also reported reliability (n = 80). Few studies reported evidence related to consequences of testing and bias (n = 4) response processes (n = 6; see Table 3).

The measures found predominantly examined teachers' beliefs or attitudes about teaching mathematics (e.g., teachers' self-efficacy for teaching mathematics) or the teaching practices they use. Measures of teacher noticing were also fairly common.

Of the instruments we located through this search thus far, only seven explicitly connect to issues of equity:

- Fennema-Sherman Mathematics Attitude Scale (Fennema & Sherman, 1976)
- Mathematics as a Gendered Domain (Forgasz et al., 2004)
- Who and Mathematics (Forgasz & Leder, 2001)
- Caring teaching practices in multiethnic mathematics classrooms (Averill, 2012)
- Views about how to support struggling students (VSSS; Jackson & Gibbons, 2014)
- Mathematics Teachers' Beliefs About English Language Learners Survey (MTBELL; Gann et al., 2016)
- Equity Quantified in Participation (EQUIP; Reinholz & Shah, 2018)

Many other instruments were not designed specifically for examining issues of equity, but have been used to study equity nonetheless. For instance Franco et al. (2007) examined differences in reform-oriented teaching related to student achievement and whether this type of teaching could help to minimize the achievement gap between schools of high and low socioeconomic status in Brazil. This small number of equity-oriented measures does not mean that these are the only measures that exist in mathematics education, but these were the only ones we found which examine teachers' affect and behavior, described in the 24 journals that we searched, from 2000-2020.

Discussion

Cultivating equitable education systems in the long-term will require policy changes at the local and national levels. Given the continued focus on randomized control trials and large quantitative studies as the "gold standard," mathematics education researchers hoping to influence policy must engage in rigorous quantitative research. Our study indicates that as a field we need to work together to create measures that are focused on issues of (and constructs related to) equity and that have validity and reliability evidence. Additionally, we found a distressing lack of studies that collected evidence of consequences of testing and bias. In order to consider equity issues related to a measure, researchers must investigate this form of validity evidence as well as test content and response processes.

We call on journal editors to expand their aims to include validation studies, as well as to press authors to use measures with established validity and reliability evidence. We acknowledge there are mathematics education instruments in the domains of teacher behavior and affect

published outside the 24 journals that we searched. We encourage authors who have developed measures to submit those measures to our cataloged inventory for review and use by others.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. (#1920621; #1920619). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Alexander, P. A. (2020). Methodological guidance paper: The art and science of quality systematic reviews. *Review of Educational Research, 90*(1), 6-23.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

Boston, M., Bostic, J., Lesseig, K., Sherman, M. (2015). Classroom Observation tools to support the work of mathematics teacher educators. Invited manuscript for *Mathematics Teacher Educator, 3,* 154-175.

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., Carney, M., & Shih, J. (2019). Reflecting on the past and thinking ahead in the measurement of students' outcomes. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 205-229). New York, NY: Routledge.

Carney, M., Bostic, J., Krupa, E., & Shih, J. (in press). Instruments and use statements for instruments in mathematics education. *Journal for Research in Mathematics Education*. Accepted for publication.

Averill, R. (2012). Caring teaching practices in multiethnic mathematics classrooms: Attending to health and well-being. *Mathematics Education Research Journal*, *24*(2), 105-128.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Sage.

Fennema, E., & Sherman, J. (1976). Fennema-Sherman mathematics attitude scales. *JSAS: Catalog of selected documents in psychology, 6*(1), 31.

Forgasz, H., & Leder, G. (2001). "A+ for girls, B for boys": Changing perspectives on gender equity and mathematics. In W. Atweh, H. Forgasz, & B. Nebres (Eds.), *Sociocultural research on mathematics education: An international perspective* (pp. 347-366). Erlbaum.

Forgasz, H. J, Leder, G. C., & Kloosterman, P. (2004). New perspectives on the gender stereotyping of mathematics. *Mathematical Thinking and Learning, 6*, 389-420.

Franco, C., Sztajn, P., & Ortigão, M. I. R. (2007). Mathematics teachers, reform, and equity: results from the Brazilian National Assessment. *Journal for Research in Mathematics Education*, *38*(4), 393-419.

Gann, L., Bonner, E. P., & Moseley, C. (2016), Development and Validation of the Mathematics Teachers' Beliefs About English Language Learners Survey (MTBELL). *School Science and Mathematics, 116*: 83–94. doi: 10.1111/ssm.12157

Jackson, K. J., & Gibbons, L. (2014, April). Accounting for how practitioners frame a common problem of practice—Students' struggle in mathematics. In D. L. Ball (Discussant), *Exploring relations between teachers' knowledge, perspectives, and practice* [symposium]. National Council of Teachers of Mathematics Research Conference, New Orleans, LA.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.2307/23353796

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. https://doi.org/10.1037/0003-066X.50.9.741

Reinholz, D. L., & Shah, N. (2018). Equity analytics: A methodological approach for quantifying participation patterns in mathematics classroom discourse. *Journal for Research in Mathematics Education*, *49*(2), 140-177.

Thunder, K., & Berry, R. Q. (2016). Research commentary: The promise of qualitative metasynthesis for mathematics education. *Journal for Research in Mathematics Education, 47*(4), 318-337.

U.S. Department of Education (2008), *What Works Clearinghouse, Procedures and Standards Handbook Version 2*, Washington: DC, December

Ziebarth, S., Fonger, N., & Kratky, J. (2014). Instruments for studying the enacted mathematics curriculum. In D. Thompson, & Z. Usiskin (Eds.), *Enacted mathematics curriculum: A conceptual framework and needs* (pp. 97-120). Charlotte, NC: Information Age Publishing.

Table 1.
Description of Five Sources of Validity Evidence

Source of Validity Evidence	Description
Test Content	"Test content refers to the themes, wording, and format of the items, tasks, or questions on a test" (AERA et al., 2014, p. 14).
Response Processes	**"**Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers." (AERA et al., 2014, p. 15)
Relationship to Other Variables	Relations to other variables may provide evidence, for example, that indicates how "...test scores [may or may not be] influenced by ancillary variables such as [individual or group characteristic]" (AERA et al., 2014, p.12).
Internal Structure	"Analyses of the internal structure of a test can indicate the degree to which the relationships among the items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p.16).
Consequences of Testing and Bias	"...decisions about test use are appropriately informed by validity evidence about intended test score interpretations for a given use, by evidence evaluating additional claims about consequences of test use that do not follow directly from test score interpretations, and by value judgments about unintended positive and negative consequences of test use." (AERA et al., 2014, p. 21)
Reliability	"The term reliability has been used in two ways in the measurement literature. First, the term has been used to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effort on performance on the second form. Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported." (AERA et al., 2014, p. 33)

Table 2.
Named Instruments Measuring Teachers' Behavior and Affect

Types of Instruments	**Number Found**
Criterion-referenced	7
Diagnostic	1
Formative	2
Interview	7
Likert/Rating Scale	5
Observation	22
Rubric	6
Survey	83

Table 3

Frequency of Validity Evidence Reported

	Number Found
Source of Validity Evidence	
Test Content	40
Response Processes	6
Relationship to Other Variables	27
Internal Structure	33
Consequences of Testing and Bias	4
Reliability	80