An Exploration of Undergraduate Mathematics Education Measures and Their Validity Evidence

Christine K. Austin
Texas State University

Kate Melhuish
Texas State University

Jim Gleason
The University of Alabama

Yvonne Lai
University of Nebraska-Lincoln

This report discusses the validity evidence reported in undergraduate mathematics education measures. Through a comprehensive literature review of articles published from 2000-2019, we identified 166 measures that provided validity claims and evidence to varying degrees. Findings overall suggest that validity evidence can be more robust within measures of RUME constructs, although there are examples of rigorous validity arguments. Of the validity evidence, the major sources of validity were reliability, test content, and internal structure. We report on the number of sources of validity, by construct, and the types of evidence to support the validity claims.

Keywords: validity evidence, undergraduate, measures, reliability

Measures play a key role in research in undergraduate mathematics education. Measures provides a means to quantify elements of teaching and learning to investigate quantitative hypotheses. Types of measures might include cognitive measures (e.g., Precalculus Concept Assessment; Carlson et al., 2010), affective measures (e.g., Motivated Strategies for Learning Questionnaire; Pintrich & de Groot, 1990), or learning environment measure (e.g., Inquiry-Oriented Instructional Measure; Kuster et al., 2019) As researchers, we have observed the use of many different measures in our field, but it is less clear the degree to which they have sufficient validity evidence to assure that scores and interpretations of them are meaningful. Previous literature has shown a lack of sufficient validity and reliability evidence in mathematics education assessments, broadly (Bostic et al., 2019; Krupa et al., 2019). Thus, we aim to explore whether this claim holds true in undergraduate mathematics education research. This report addresses:

RQ1: What measures are currently in use by RUME researchers?

RQ2: What types of validity evidence support the validity of their usage?

Background

Measurement theory is the process of investigating varying forms of evidence to form the basis of a valid and reliable instrument (Crocker, 2012). For claims of an instrument to be meaningful the instrument needs to be assessed to ensure it is measuring what it was intended to. In doing so, researchers assess the reliability and the validity of the instrument. Reliability is the degree in which the results/scores of an instrument are replicable, consistent, and accurate (Brennan, 2006; Crocker & Algina, 2006). Some examples for evidence of reliability include test-retest, alternate test forms, item response theory, Cronbach alpha, and inter-rater reliability. Validation, on the other hand, has been described by Cronbach (1971) as the collection of evidence to support the claims drawn from the results of the instrument by the assessment creator or the user of the instrument. In education, validity is defined in the *Standards for Educational and Psychological Testing* as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (American Educational Research Association [AERA] et al., 2014, p. 11). In the *Standards*, five forms of validity are noted: test content, response process, internal structure, relation to other variables, and consequence of testing (AERA et al., 2014). Primary sources of evidence for the five types of validity are provided in the *Standards* (AERA at al., 2014) and *Validation in Mathematics Education* (Krupa et al., 2019).

Five Sources of Validity

Test content validity is explained as "the relationship between the content of a test and the construct it is intended to measure" (AERA et al., 2014, p. 14). Evidence for test content stems from the judgments of experts, observation, and logical or empirical analyses (AERA et al., 2014, p. 14). In practice, such validity evidence includes, but not limited to, data from experts (i.e., experts in the field), field work, literature reviews, alignment of a learning trajectory, and participant generated content (i.e., the participant creating the data for the researcher) (Krupa et al., 2019). *Response process* validity is concerned with "the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers" (AERA et al., p. 15). In other words, this form of validity ensures that the test takers engage in the appropriate thinking, processes, and analyses in which the construct was designed to be measured for proper claims to be made (Lavery et al., 2019). A non-exhaustive list of Validity evidence for response processes includes, cognitive interviews, focus groups, eye-tracking, predicted response patterns based on learning trajectories, and think alouds (AERA et al., 2014, Krupa et al., 2019; Lavery et al., 2019).

Validity from *internal structure* addresses "the degree to which the relationships among test items and test components conform to the construct" (AERA et al., 2014, p. 16). This validity type is focused on confirming that the construct adheres to the theories in which it was built upon (Lavery et al., 2019; Krupa et al, 2019). Further, validity evidence of internal structure confirms that the construct is unidimensional. A unidimensional construct assesses a single dimension or attribute (e.g., a student's level of multiplicative reasoning) (Bond & Fox, 2015). Validity evidence of internal structure includes, but is not limited to, factor analysis, Item Response Theory, Rasch modeling, and multidimensional scaling (AERA et al., 2014, Krupa et al., 2019; Lavery et al., 2019). *Relation to other variables* validity is defined as "the degree to which the relationships among test items and test components conform to the construct" (AERA et al., 2014, p. 16). Many instruments are developed with the expectations that the construct will relate to other constructs or variables in predictable ways or patterns (Lavery et al., 2019) thus, this source of validity evidence connects results to other constructs (Krupa et al., 2019). Such types of validity evidence include correlations analysis, regression analysis, structural equation modeling, hierarchical linear modeling, and alignment with expert opinion of user (e.g., teacher). Lastly, validity of *consequences of testing* is the process of "gathering evidence to evaluate the soundness of [the] proposed interpretations for their intended use" (AERA et al., 2014, p. 19). *Standards for Educational and Psychological Testing* discuss three main types of consequences of testing: unintended consequences (e.g., differences in test score based on race/ethnicity), claims made that are not based on test score interpretations (i.e., the need for new more information to address the results of the testing), and the interpretations and uses of the test scores intended by the instrument creators (AERA et al., 2014). Evidence for consequence of testing validity include explicit intended use/interpretations, motivational consequences, differential item functioning, and appropriate cut scores (AERA et al., 2014; Krupa et al., 2019).

Methods

This project began in 2020 as part of an NSF-funded collaboration (BLINDED) to catalog mathematics education measures that have been in use for the last twenty years. The authors of this paper have led the work on the undergraduate sub-team. To identify measures in use, we conducted a literature search spanning 2000-2019. We searched the mathematics education research journals identified by Williams and Leatham (2017) along with *International Journal of Research in Undergraduate Mathematics Education* as its newness means it was not contained in Williams and Leatham's list. We used the following Boolean string to identify articles: (tertiary OR post-secondary OR undergraduate OR college OR university OR "graduate student") AND (test* OR assess* OR instrument* OR measur* OR survey* OR inventor*). From this initial search, we identified 2,603 articles. At this point, we developed an inclusion and exclusion system with for criteria: (0) Must focus on undergraduate or graduate mathematics students and/or instructors (exclusive of primarily statistics or pre-service teachers as they fall under other subteams (1) Must be an empirical study (2) Used an instrument that is uniform across the population, that should be useable by other researchers with at most minimal modification, potentially in a different setting (3) That is associated with a named or described construct. After training with a set of 30 articles within the full set, the remaining articles within the 2,603 articles were coded according to these criteria by two members of the research independently, then coming to agreement about whether to include to the next stage. Here, 432 articles met these criteria. At this point, we switched to identifying measures. To be considered a measure, we required a composite score of some type. That is, an instrument that only reports data at an individual item level was not included. We identified 241 unique measures.

The final stage of this process was then to identify validity evidence. For each measure, we went into the article where it was found, then determined if that article contained its validity evidence. If not, we searched for another source of evidence (such as a reference validation study.) For any measure with validity evidence, we documented: focal construct (cognitive, affective, learning environment), usage information, and any validity evidence or claims provided categorized within the five types of validity evidence in the prior section. We aligned our classifications with the larger project teams' set of evidence types. In the next section, we present an overview of the types of instruments and most common types of validity evidence.

Results

Of the 241 unique measures identified to fit our criteria, only 166 measures were reported to have validity evidence for the constructs. Ultimately, 75 measures were eliminated due to having no apparent reported validity evidence within the manuscript. Our final list of 166 undergraduate mathematics articles contained three prominent construct types that were broken up into their focal construct: affective domain, cognitive domain, and learning environment (see Table 1). Our findings show that there is a skew in the type of constructs being implemented in undergraduate research. These results indicate that the majority of the instruments measure the affective (47%) and cognitive domain (45%) and only about a tenth (11%) are designed to measure an undergraduate student learning environment.

On average, two (M = 1.92, SD = 1.06) forms of validity evidence (including reliability) were reported for each measure. Only one measure provided all sources of validity and reliability evidence. Given that most of the articles only contained two sources of validity, on average, the most prevalent source of validity was computed. Table 1 illustrates the validity evidence broken down by construct type. The most reported form of validation was test content (n = 113) followed by reliability (n = 87). Very few undergraduate mathematics measures reported validity evidence for consequences of testing (n = 4) or response process (n =15).

Table 1. Source of Validity by Construct Type

Source of Validity	Construct Type [Total # of Articles]							
	AD [72]	AD / CD [2]	CD [72]	CD / LE [1]	LE [13]	LE / AD [4]	Other [2]	Total [166]
Reliability	46	1	25	0	9	4	2	87
Test Content	43	1	56	1	8	0	1	110
Response Process	3	0	9	0	3	0	0	15
Internal Structure	36	1	16	0	9	1	1	64
Relation to Other Variables	18	0	15	0	4	0	1	38
Consequence of Testing	0	0	4	0	0	0	0	4

Note: AD = Affective Domain, CD = Cognitive Domain, and LE = Learning Environments

Table 2 reports the types of validity evidence that were documented for the 166 measures. Our findings indicate that of our 87 measures that made claims for reliability, 72 (83 %) reported a Cronbach Alpha Coefficient as the source of evidence. Further, the 38 measures that provide evidence for relation to other variables predominantly used correlation analysis (n = 27) as their source of evidence (71%). As shown in Table 2, other than evidence type for test content, there is little variation to the evidence type documented undergraduate mathematics education research.

While in general, the validity evidence for measures was scarce, there were some notable exceptions in the set. Of the affective focused measures, *Mini-IPIP Scales* (Alcock et al., 2014) and *Attitudes to Technology in Mathematics Leaning Questionnaire* (Forgarty et al., 2001) had 7 and 6, respectively, forms of validity evidence (see Table 2 for validity evidence types). The cognitive focused measures with 6 or more types of validity evidence consisted of the *Group Theory Concept Assessment* (Melhuish, 2019), *Calculus Readiness Assessment* (Pyzdrowski et al., 2013), *Function Concept Inventory* (O'Shea et al., 2016), *College Readiness Assessment* (Bernbaum et al., 2011), *Calculus Concept Inventory* (Bison et al., 2016), and *Proof Comprehension Tests* (Mejía-Ramos et al., 2017). Lastly, of learning environment focused measures, *Survey of Calculator Usage Extensiveness and Subordinality* (Mao et al., 2017), *Concept Test* (Kaw & Yalcin, 2012), and *Inquiry-Oriented Instructional Measure* (Kuster et al., 2019). While we highlight these measures, we note that it is not the number of validity evidence types that matters so much as the integration of validity evidence into a validity argument. Nonetheless, when validity evidence is presented primarily as reliability, it is difficult to marshal this evidence into a robust argument that an instrument measures a particular construct as intended.

Table 2. Validity Evidence

Source	Evidence Type	
Reliability	Alternate Test Forms	1
	Inter-Rater Reliability (Kappa & Percent Agreement)	7
	Cronbach Alpha	72
	Item Response Theory and/or Rasch Modeling	2
	Item Remainder Correlations	2
	Kuder-Richardson Formula 20	4
	Sensitivity Analysis	1
	Split-Half Reliability	2
	Test-Retest	10
Test Content	Alignment with Framework/Theory/Learning	24
	Data from Experts	38
	Fairness of Content	2
	Field Work	4
	Literature Review	36
	Participant Generated Content	16
	Revision Process	15
	Standard Setting	1
Response Process	Cognitive Interviews	4
	Rater Agreement/Reliability	3
	Rater Training and Collaboration	2
	Student Written Work	5
	Think Aloud	2
Internal Structure	Factor Analysis*	58
	Item Difficulty	4
	Item Scale Correlations	8
	Item Response Theory (IRT)	6
	Rasch Modeling	5
Relation to Other Variables	Alignment with Expert Opinion	1
	Convergent or Divergent Association	4
	Correlation Analysis	27
	Discriminant Validity	1
	Hierarchical Linear Modeling (HLM)	1
	Statistical Testing (e.g., regression, t-test, chi-square)	5
	Triangulation	1
Consequence of Testing	Appropriate Cut Score	2
	Cost-Benefit Analysis	1
	Item Function (e.g., DIF)	1

*Includes Bifactor, Exploratory/Confirmatory Factor Analysis, Parallel Analysis, Principal Axis Factoring, and Principal Component Analysis

Discussion

Our analysis of measures used in the undergraduate mathematics education literature points to several findings. First, we found that Bostic et al.'s (2019) and Krupa et al.'s (2019) reflection on the state of measures in math education was largely true in the undergraduate context. While many instruments are in use many provided limited or no validity evidence at all. This is particularly problematic when scores from such measures are used in service of testing particular hypothesis or have impacts on students or instructors in practice. That is, the trustworthiness of an instrument is determined by the degree we can be assured it is valid.

If we consider our results more broadly, we can see that certain types of validity evidence are more frequently reported. For example, Cronbach's alpha is by far the most common type of validity evidence in our data set. This is not surprising as it is a relatively easy type of statistic to calculate with associated indices of what is acceptable for different measure purposes. We also documented other common psychometric and quantitative approaches to validity with the prevalence of factor analysis and correlation with variables. These common approaches still accounted for less than half of the measures reporting any validity evidence.

We note the measures were more frequently reported with quantitative evidence and less qualitative validity evidence. We suggest a robust validity argument (e.g., Kane, 1992) is made with mixed methods. For example, response process evidence was quite infrequent. This type of evidence stems from making sure that participant response to items actually reflects their thinking. This can be done through cognitive interviews, collecting written work, and other approaches. Without such evidence, we may be left wondering the degree to which an instruments' scores truly reflect participants.

We acknowledge some limitations of the current work. First, our exclusion and inclusion process mean that we may have overlooked instruments that have not been discussed in primary mathematics education journals. Additionally, it is possible that we overlooked a source of validity evidence that might be found in harder to access internal reports. There is also likely some variation in how our coders interpreted different types of evidence and therefore, we suggest caution when generalizing from frequencies that are similar.

Finally, we note directions for future research. In our own data set, we have made some initial notices that within our construct categories we can find some interesting patterns. For example, a full 28 of the affective instruments were focused on mathematics attitudes and belief scales. This means that there are many instruments in use targeting the same construct. We also note that within the cognitive construct measures, instruments that aligned themselves with concept inventory procedures (e.g., Group Theory Concept Assessment, Function Concept Instrument, Calculus Concept Inventory) often had more robust amounts of validity evidence likely due to commonalities in development approach. The next research direction will involve better understanding the nature of the instruments within each category and their relation to the types of validity evidence shared.

We conclude by suggesting that the field consider a variety of validity evidence types when creating measures. It may be a truism that the validation process is never complete. Each new sample and use require recalibration. The stronger the evidence is for our measures, the more we can trust the results of our quantitative studies.

References

Alcock, L., Attridge, N., Kenny, S., & Inglis, M. (2014). Achievement and behaviour in undergraduate mathematics: personality is a better predictor than gender. *Research in Mathematics Education, 16*(1), 1-17.

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

Bernbaum Wilmot, D., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a learning progression in mathematical functions for college readiness. *Mathematical Thinking and Learning*, *13*(4), 259-291.

Bisson, M. J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*, *2*, 141-164.

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

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*, *28*(2), 113-145.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulleti*n, 112(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kuster, G., Johnson, E., Rupnow, R., & Wilhelm, A. G. (2019). The inquiry-oriented

Crocker, L., & Algina, J. (2006) *Introduction to classical and modern test theory*. Thomson Wadsworth.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.). *Educational Measurement* (2nd Ed.). Washington, D.C.: American Council on Education.

Fogarty, G., Cretchley, P., Harman, C., Ellerton, N., & Konki, N. (2001). Validation of a questionnaire to measure mathematics confidence, computer confidence, and attitudes towards the use of technology for learning mathematics. *Mathematics Education Research Journal*, *13*, 154-160.

Kaw, A., & Yalcin, A. (2012). Measuring student learning using initial and final concept test in an STEM course. *International Journal of Mathematical Education in Science and Technology*, *43*(4), 435-448.

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

Kuster, G., Johnson, E., Rupnow, R., & Wilhelm, A. G. (2019). The inquiry-oriented instructional measure. *International Journal of Research in Undergraduate Mathematics Education*, *5*, 183-204.

Lavery, M, R., Jong, C., Krupa, E. E., & Bostic, J. D. (2019). Developing an instrument with validity in mind. In J. D. Bostic, E. E. Krpua, & J. C. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 12-39). Routledge.

Mao, Y., White, T., Sadler, P. M., & Sonnert, G. (2017). The association of precollege use of calculators with student performance in college calculus. *Educational Studies in Mathematics*, *94*, 69-83.

Mejía-Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education*, *19*(2), 130-146.

Melhuish, K. (2019). The Group Theory Concept Assessment: A tool for measuring conceptual understanding in introductory group theory. *International Journal of Research in Undergraduate Mathematics Education*, *5*(3), 359-393.

O'Shea, A., Breen, S., & Jaworski, B. (2016). The development of a function concept inventory. *International Journal of Research in Undergraduate Mathematics Education, 2,* 279-296.
Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33–40. https://doi.org/10.1037/0022-0663.82.1.33
Pyzdrowski, L. J., Sun, Y., Curtis, R., Miller, D., Winn, G., & Hensel, R. A. (2013). Readiness and attitudes as indicators for success in college calculus. *International Journal of Science and Mathematics Education, 11*, 529-554.
Williams, S. R., & Leatham, K. R. (2017). Journal quality in mathematics education. *Journal for Research in Mathematics Education*, *48*(4), 369-396.