Session Title: Argument-Based Approaches to Measure Validation Situated in Mathematics Education

Discussant: Mark Wilson

Speaker: Michele Carney

Abstract: This symposium builds on the work from the the National Science Foundation funded the conference *Validity Evidence for Measurement in Mathematics Education* (V-M^{2}Ed). We will first present a review of the literature on argument-based approaches to measure validation, followed by five examples of argument-based validation studies for instruments used in mathematics education research and practice. Together, the five papers (1) provide models for the use of arguments as a measurement validation methodology, (2) highlight affordances and constraints of argument-based approaches to validation, and (3) posit suggestions for structuring arguments. Following the five papers, the discussant will facilitate discussion between the panelists and audience on arguments as a measurement validation methodology.

Session Summary:

The Standards for Educational and Psychological Testing (2014) promote an argument-based approach to validation.

> Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretations of test scores and their relevance to the proposed uses (p. 11)...Decisions about what types of evidence are important for the validation argument in each instance can be clarified by developing a set of propositions or claims that support the proposed interpretation (p. 12).

Michael Kane (1992) put forth an approach for structuring arguments that develop a chain of reasoning using a set of propositions or claims from an observed performance to a score interpretation. Schilling and Hill (2007) modified Kane's approach, suggesting a more prescriptive structure that delineates common assumptions and inferences would provide structure to instrument developers and ensure important considerations were not missed. However, neither approach appears to be widely used in practice. The Standards for Educational and Psychological Testing (2014) delineate five sources of validity evidence (i.e., test content, response process, internal structure, relation to other variables, and consequences). While these sources can be used to identify common assumptions and inferences, a clear framework for structuring the argument is absent.

In the spring of 2017, the National Science Foundation funded the conference *Validity Evidence for Measurement in Mathematics Education* (V-M^{2}Ed). A key outcome for the conference was to contextualize argument-based approaches to validation within the field of mathematics education. A group of researchers from the conference are working together to articulate structures for validation arguments while also providing examples of these arguments. The goal of this work is to foster discussion and reflection within our fields to improve the quality of mathematics education instruments and validation methodology.

Our symposia session will present the work of these researchers. The first paper presents a review of the literature on validation methodology with a specific focus on argument-based approaches in mathematics education. Papers 2-5 present validation arguments for particular instruments used within mathematics education while also specifying and justifying their specific validation approach. Paper 6 will compare and contrast two argument-based approaches for one instrument. Structured time for discussion with audience participants will follow. Our objectives are to:

1. Provide examples for the use of arguments as a measurement validation methodology,
2. Highlight affordances and constraints of argument-based approaches to validation, and
3. Posit suggestions for structuring arguments

By situating the validation argument discussion within the content area of mathematics education we hope to further foster conversation between mathematics education and measurement researchers.

This work is significant because while there are multiple discussions in the literature about validity (e.g., see issue 2 in 2016 of Assessment in Education: Principles, Policy & Practice) there are a few empirical examples of argument-based approaches to validation. This is particularly true within the area of mathematics education. We need consistent methodology around establishing validity of score interpretations for proposed uses. This symposia will assist by providing examples and fostering discussion in these areas.

Paper #1

Title: Examining the Arguments Surrounding the Argument-Based Approach to Validation: A Systematic Review of Validation Methodology

Presenting Author: Matthew Lavery

Non-Presenting Authors: Michele Carney, Jonathan Bostic, Jeff Shih, Erin Krupa, Mark Wilson, Lance Kruse

Paper/Presentation Summary:

As early as Descartes (1637/1970), logic and reason have been positioned as tools for individuals to advance their own understanding. By contrast, argumentation is an interactive, social exercise used for persuasion, collective cognition, and to advance shared knowledge (Mercier & Sperber, 2011, 2017). When one advances an argument, subjects it to the tests and challenges of others, and responds to questions and counterarguments, one's thinking improves (Mercier & Sperber, 2017). Through argumentation, groups produce correct solutions more often than individuals (Moshman & Geil, 1998) and individual accuracy improves as well (Castelain, Girotto, Jamet, & Mercier, 2016). Since it was formally introduced by Kane (1990, 1992), the argument-based approach to validation has been promoted in the field of educational and psychological measurement as the preferred method for validating interpretations and uses of test scores (AERA, APA, & NCME, 2014; Kane, 2013; Schilling & Hill, 2007). Scholars continue to debate the best approaches for developing and supporting validity arguments, however (for examples, see Brennan, 2013; Kane, 2007).

Purpose and Perspective

Since validation is currently discussed in terms of arguments, and since arguments are both interactive and social, the purpose of the present review is to systematically examine the structure and content of the scholarly arguments about validity and validation which appear in the peer-reviewed literature. Using theories of argumentation as a lens, researchers examine the arguments and counterarguments offered in the literatures reviewed to identify key assertions and recommendations regarding validity arguments and validation methodology. Researchers then analyze the validity arguments and evidence reported in peer reviewed journals on specific interpretations and uses of test scores to determine the degree of alignment between validation theory and practice.

Method and Data Sources

Researchers used the EBSCOhost platform to search the Education Full Text (H.W. Wilson), Education Research Complete, ERIC, and PsycINFO databases for articles published in peer-reviewed journals within the past 15 years that contain either the words "validity argument", or the words "argument-based approach" along with "validity" or "validation". After duplicates were removed, the search returned n = 168 articles. Per the PRISMA statement (Moher, Liberati, Tetzlaff, & Altman, 2009), researchers examined titles and abstracts to further qualify articles for the study. Articles which discuss and make recommendations regarding validity arguments qualified for inclusion in the study as methodological publications (n = 83 articles). Articles which present interpretation arguments, score-use arguments, and/or validity arguments for specific tests, along with validity evidence to support (or challenge) those arguments qualified for inclusion in the study as applied articles (n = 85 articles). Ten articles qualified in both categories, while another 10 articles were excluded from the study.

Findings and Significance

While the literatures reviewed identify a few competing concerns about validity arguments, methodological papers consistently caution against collection and reporting of validity evidence that does not directly support a thoughtful, context-specific argument. By contrast, applied papers demonstrate only partial satisfaction of methodological recommendations. The findings of this review provide a framework for the integration of methodological recommendations into an accessible framework for applied researchers.

Paper #2

Title: A Validity Argument for an Innovative Assessment System based on Learning Trajectories

Presenting Author: Jere Confrey

Non-Presenting Authors: Garron Gianopulos, Meetal Jaswant Shah

Paper/Presentation Summary:

We report on what we have learned in our efforts to build a validity argument for assessments embedded in a digital learning system (DLS) for middle grades mathematics (Author, 2015). Our validation approach borrows from Kane's interpretive argument (2004), argument mapping (Wigmore, 1913; Toulmin, 1958), and Popper's concepts of falsifiability and auxiliary theories (1962). We also integrated guidelines from the standards for educational and psychological testing (2014) and the CCSSO's "Criteria for High-Quality Assessment" (2014).

Our validation work was performed on the assessments within Math-Mapper 6-8 DLS. These diagnostic assessments were built around Learning Trajectories (LT). LTs document landmarks and obstacles that students may encounter as they progress from a naïve to sophisticated understanding of a target concept (Confrey, Maloney, & Corley, 2014).

We designed our system of assessments with four objectives in mind: Firstly, score reports will provide actionable and accurate student- and class-level feedback so that teachers can plan and inform instruction in a theory-driven manner. Secondly, if teachers interpret reports according to our guidelines, they will draw valid conclusions concerning the progress of students. Thirdly, students will know what they understand more precisely, and see a clear path to improve. Fourthly, if teachers use the conclusions to adapt instruction, learning gaps will close, misconceptions will diminish, and overall learning will increase.

Given the central role LTs play in our system and score reporting, the internal structure of the tests are a critical element of our validity argument. Therefore, we made two predictions with respect to the internal structure: LT items would be essentially unidimensional, and LT levels would positively correlate with item difficulty. To test these predictions, we conducted exploratory and confirmatory factor analyses, examined scatter plots of LT level and Rasch item difficulty (Wilson, 2005).

We have collected two years of field test data from two school districts. Sample sizes ranged from 200 to 2000 responses per test. The predicted correlation between LT level and item difficulty did surface in the majority, but not all of the LTs. We will present our interpretation of these findings in light of Popper's (1962) concept of auxiliary theories. This presentation will have scholarly and scientific significance because it exemplifies one approach to integrating a validity argument into test development with the goal of improving the quality and coherence of validation arguments.

Paper #3

Title: Measuring Knowledge and Motivation for Teaching Multidigit Arithmetic: Evidence of Elemental, Structural, and Ecological Validity

Presenting Author: Erik Jacobson

Non-Presenting Authors: Dubravka Svetina

Paper/Presentation Summary:

Purpose

The mathematical proficiency for teaching framework (Author, 2013) identifies a multifaceted goal for elementary preservice teacher education: integrated knowledge and productive disposition for teaching. Research on how teacher education influences mathematical proficiency for teaching is difficult because existing measures differ in focus and scope. The purpose of the study was to develop a coordinated measure of knowledge and motivation for teaching multidigit arithmetic. We report validity evidence for the novel measure.

Framework

In an argument-based approach to validity, various forms of evidence are used to support argument-based inferences regarding score interpretation and use. Specifically, we address three aspects of validity: elemental, structural, and ecological (Schilling & Hill, 2007). The elemental aspect concerns the items (i.e., content validity); the structural aspect concerns how items are combined in subscales; and the ecological aspect concerns how the measure is related to the context of use. For each aspect, we discuss theory-based assumptions, articulate inferences based on these assumptions, and assess empirical evidence to support the inferences.

Methods, Data, and Results

Elemental. We assumed the items reflected the target constructs, not extraneous factors. We inferred that knowledge would correspond with item responses, and used 60-minute think-aloud interviews (n = 15) to confirm. We also inferred that experts would judge the knowledge and motivation items to be consistent with the construct, and surveyed 8 experts to assess this inference.

Structural. We assumed items would form unidimensional scales for each construct, and furthermore that the motivation items would be more closely related with each other than with the knowledge items. Therefore, we inferred that the items for each scale were unidimensional. We also inferred that a model with the hypothesized structure would better fit the data than competing models. We used data from two survey administrations (n = 169) to confirm these inferences.

Ecological. We assumed that the measure was sensitive teacher education and appropriately related to other constructs. Thus, we inferred PSTs knowledge and motivation would increase during a methods class, and used pre-post surveys of 33 PSTs to confirm. We also inferred that knowledge was correlated with conceptions of multidigit number (Thanheiser, 2009) and that motivation was correlated with teaching self-efficacy beliefs (Tschannen-Moran & Hoy, 2001), and used survey data (n = 114) to confirm.

Significance

The validity evidence for the novel measure suggests that it accurately reflects the intended constructs for the intended use. Thus, continued use of the measure to investigate teacher education within the MPT framework is justified.

Paper #4

Title: Instantiating the validity argument framework: Evaluating the validity of the uses of universal screeners

Presenting Author: Leanne Ketterlin-Geller

Paper/Presentation Summary:

Objectives: Despite the call for an argument-based approach to validity over 25 years ago, few examples exist in the published literature. The purpose of this manuscript is to illustrate the argument-based approach to validity for the uses and interpretations of a universal screener for middle-school mathematics. Specifically, a universal screening assessment system was created for use within a RtI framework to help middle school teachers support students' algebra readiness. This universal screener is a computer-based multiple-choice assessment that takes approximately 20 minutes to complete, and is administered three times per year (fall, winter, and early spring).

Theoretical Framework: Universal screeners are formative assessments that form an integral part of a comprehensive assessment framework for implementing Response to Intervention (RtI). Results from universal screeners help teachers make instructional decisions early in the learning process to prevent and remediate skill gaps. Data are intended to help teachers (1) identify students who are at-risk for future failure in the domain and then (2) determine the level of intensity of supplemental instructional support that may help at-risk students reach their learning goals. Because of the important use of universal screener results, sufficient validity evidence should substantiate the trustworthiness and meaningfulness of the results from making these decisions.

A validity argument is the process of creating an evidence-based case for the intended interpretations and uses of the observed score. Messick (1995) summarized the importance of this process, stating, "score validation is empirical evaluation of the meaning and consequence of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretations and use" (p. 5). The extent to which a score-based decision is justifiable depends on the clarity and coherence of the accumulated evidence (Kane, 1992, 2006, 2013).

Modes of Inquiry: First, this paper presents an interpretive and use argument (IUA) that outlines the inferences and assumptions leading from the score on the universal screener to the intended uses and interpretations. The inferences are sequenced following Kane's (1992, 2006) structure to progress from scoring to generalization to extrapolation inferences. The assumptions underlying each inference serve as the testable hypotheses from which evidence is collected in a validity argument.

Data Sources: Second, this paper presents a validity argument by gathering evidence designed to address the hypotheses outlined in the IUA. Evidence is derived from procedural sources collected during the test development process, psychometric analyses from a large-scale field test, descriptive and correlational studies conducted during pilot testing, and interviews and observations of teachers using the results to make classroom-based decisions.

Substantiated Conclusions and Implications: The defensibility of the evidence is evaluated and an overall evaluation of the validity of the interpretations and uses of the universal screener is provided. Implications for practitioners using the universal screener are proposed, and the findings are generalized to inform test developers undergoing similar efforts.

Paper #5

Title: Affordances and Constraints of Two Approaches to Validation Arguments

Presenting Author: Michele Carney

Non-Presenting Authors: Carl Siebert, Keith Thiede, Angela Crawford, Richard Osguthorpe

Paper/Presentation Summary:

In spring of 2017 measurement and mathematics education researchers came together to examine argument-based approaches to measure validation within the context of mathematics education. Several frameworks were presented as potential approaches with a focus on articulating the claims and assumptions within validation arguments. The purpose of this paper is to compare and contrast two of these frameworks by articulating two different validation arguments for the same instrument. Our goal is to identify the affordances and constraints provided these two frameworks.

The two frameworks are:

- The interpretative argument aspect of Kane's (2004) *observed performance to interpretation for use* approach the first step of which is stating the chain of assumptions and inferences that start at the process of converting an observed performance to a score for an item and ends at the interpretation of the test score for a particular use.
- The *sources of validity* from the Standards for Educational and Psychological Testing (2014) can also be used to identify and articulate the claims and assumptions inherent in the stated score interpretation for proposed uses.

Each framework will be used to independently articulate claims and assumption for the instrument described below.

The instrument is the Diagnostic Assessment of Proportional Reasoning (DAPR) which measures student composed unit and multiplicative comparison conceptions in proportional reasoning situations. The DAPR is a 20 item fill-in-the-blank assessment available in three equated forms and administered by classroom teachers. The assessment content is targeted at middle grades standards in the Common Core and several studies generating aspects of validity evidence using students in grades 6-9 have been conducted. A student's DAPR score can be interpreted in relation to a learning trajectory of composed unit and multiplicative comparison understanding and used by classroom teachers to identify instructional scaffolds for students.

The claims and assumptions across the two frameworks will be generated independently and then the resulting arguments compared. The goals is to identify the affordances and constraints of argument-based frameworks. For example, initial results from discussions at the *Validity Evidence for Measurement in Mathematics Education* conference indicates an affordance of Kane's framework is the press for a coherent chain of reasoning but because the argument development process differs from the typical instrument development process, it is often more difficult for developers to articulate than more traditional approaches, such as the sources of validity.

Currently, validation work in mathematics education tends to focus on presenting isolated aspects of validity evidence. However, this evidence is seldom associated with a specific claim or assumption and is rarely situated within a comprehensive validation argument. This paper can serve as an example within the mathematics education for structuring an argument that identifies specific claims and assumptions inherent in a particular score interpretation and foster discussion around the two frameworks and more generally about an argument-based approach to validation.

References

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.

Author, (2013).

Authors (2014).

Brennan, R. L. (2013). Commentary on 'Validating the Interpretations and Uses of Test Scores'. *Journal of Educational Measurement, 50*(1), 74-83. doi:10.1111/jedm.12001

Castelain, T., Girotto, V., Jamet, F., & Mercier, H. (2016). Evidence for benefits of argumentation in a Mayan indigenous population. *Evolution and Human Behavior, 37*(5), 337-342. doi:10.1016/j.evolhumbehav.2016.02.002

Descartes, R. (1637/1970). *Discours de la méthode*. New York: Liberal Arts Press.

Kane, M. T. (1990). *An argument-based approach to validation*. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED336428

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

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 135-170.

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

Kane, M. T. (2007). Validating Measures of Mathematical Knowledge for Teaching. *Measurement: Interdisciplinary Research and Perspectives, 5*(2-3), 180-187.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi:10.1111/jedm.12000

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57-74. doi:10.1017/S0140525X10000968

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge, Massachusetts: Harvard University Press.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine, 151*(4), 264-269.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning, 4*(3), 231-248. doi:10.1080/135467898394148

Schilling, S. G., & Hill, H. C. (2007). Assessing Measures of Mathematical Knowledge for Teaching: A Validity Argument Approach. *Measurement: Interdisciplinary Research and Perspectives, 5*(2-3), 70-80.

Thanheiser, E. (2009). Preservice elementary school teachers' conceptions of multidigit whole numbers. *Journal for Research in Mathematics Education, 40*(3), 251–281.

Toulmin, S. (1958). *The uses of argument*. Cambridge: UK: Cambridge University Press.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805.

Washington, DC: American Educational Research Association.
CCSSO. (2014). *Criteria for Procuring and Evaluating High-Quality Assessments*. Washington: DC.

Wigmore, John Henry (1913). *The principles of judicial proof: as given by logic, psychology, and general experience, and illustrated in judicial trials*. Boston: Little Brown.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system.
Applied measurement in education, 13(2), 181-208.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.