Argument-Based Validation in Practice: Examples from Mathematics Education

A. Erin Elizabeth Krupa (Corresponding Author)

Mathematical Sciences Department

Montclair State University

1 Normal Avenue

Montclair, NJ 07043

krupae@montclair.edu

B. Michele Carney

Curriculum, Instruction, and Foundational Studies College

1910 University Drive

Boise, ID 83725

michelecarney@boisestate.edu

C. Jonathan Bostic

School of Teaching and Learning

Bowling Green State University

529 Education Building

Bowling Green, OH 43043

bosticj@bgsu.edu

Argument-Based Validation in Practice: Examples from Mathematics Education

Authors: Erin Krupa, Michele Carney, Jonathan Bostic

Abstract

This paper provides a brief introduction to the set of four manuscripts in the special issue. To provide a foundation for the issue, key terms are defined, a brief historical overview of validity is provided, and a description of several different validation approaches used in the issue are explained. Finally, the contribution of the manuscripts to further articulating argument-based validation approaches is discussed, along with questions for the field to consider.

Keywords: Validation, validation argument, argument-based validation, mathematics education

Argument-Based Validation in Practice: Examples from Mathematics Education

Current conceptions of validity and validation, as articulated in the *Standards for Educational and Psychological Testing* (*Standards)* (American Educational Research Association [APA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 2014), focus on (a) validity as referring to the interpretation of scores for specified uses, and not the validity of a test, and (b) validation methodologies involving the presentation of an argument with supporting theoretical and empirical evidence. It is relatively common for education researchers and practitioners to refer to the validity of a test (Bostic, Krupa, Carney, & Shih, in press), and while there are examples of high quality work being conducted in validation (e.g., Mislevy and colleagues' (2003) Evidence Centered Design and Wilson's (2005) process of constructing measures), there is evidence that current conceptions of validity and validation are not widely-used within the field of education (Cizek, Rosenberg, & Koons, 2008; Shear & Zumbo, 2014; Wolming & Wikström, 2010). In particular, within the field of mathematics education there have been calls for increased reporting on the issues of validity and reliability related to instruments (Hill & Shih, 2009; Bostic 2017; Bostic, Krupa, Carney, & Shih, in press).

While there have been several journal issues related to what constitute validity (e.g., Newton, 2012; Newton & Baird, 2016), less had been done to reach consensus on or even summarize differences among approaches to the validation process. Therefore, the papers in this themed issue discuss argument-based validation frameworks and provide examples, from specific instruments in mathematics education, to illustrate their selected validity framework. The frameworks that will be described in the introduction and explored throughout this issue include: Kane's (2004, 2006) validation argument from observed performance to interpretation

for use approach, the sources of validity evidence from the S*tandards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), Schilling and Hill's (2007) modification of Kane's approach involving three common types of validation-related assumptions and inferences (i.e., elemental, structural, and ecological), the National Research Council's (2001) assessment triangle, and Pellegrino DiBello, and Goldman's (2016) framework for instructionally relevant assessments.

The intent of the articles in this issue are to provide examples that may be useful for instrument developers when thinking about validation arguments, while also providing examples from mathematics education that illustrate the strengths and challenges of various validation argument approaches. The overall goal for this issue is to fuel discussion around the examples provided in the issue to press the fields of measurement and mathematics education to further use these conceptions in practice, and to provide fodder for further discussion around argument-based approaches that are grounded in application. This article will present key terms that will be used throughout other articles in the special issue, provide a brief history of validity and an argument-based approach to validation, describe five argument-based validation frameworks, and discuss the contribution of the manuscripts to argument-based validation approaches, along with questions for the field to consider.

Defining Terms

Many researchers use the terms measure, instrument, assessment, and test interchangeably without pointing to the nuanced differences in their meaning. Since the same words are used differently by researchers and the same topics are often discussed using different terms, we want to avoid any misunderstandings. Therefore, it is important to define our conceptions of these terms, which are used consistently throughout the special issue. First, the

term *measure* will be used as a verb, to measure a construct, rather than as a noun, about a particular measuring device.

The *Standards* (AERA, APA, & NCME, 2014) define a *test* as "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" (p. 224). Similarly, Markus and Borsboom (2013) define a testing as "it covers any technique that involves systematically observing and scoring elicited responses of a person or object under some level of standardization" (p. 2). Both of these definitions incorporates the importance of having a systematic method of observation and a standardized scoring process for tests. The authors in this issue will use the term *instrument* more broadly than test to describe measures of constructs where responses are typically scaled along a continuum, such as a Likert scale for beliefs. Unlike tests, instruments are not usually related to a body of knowledge with correct and incorrect answers.

Both the *Standards* (AERA, APA, & NCME, 2014) and Markus and Borsboom (2013) consider *assessment* to be a broader term. The latter authors argue the term assessment might "include non-systematic or non-standardized methods" (2014, p. 2). The *Standards* defines assessment two ways in two different places: (1) as "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 (2014, p. 216) and (2) as "commonly referring to a process that integrates test information with information from other sources (e.g., information from other tests, inventories, and interviews; or the individual's social, educational, employment, health, or psychosocial history)" (2014, p. 2). Taken together, assessments are broader than tests and instruments because they rely on additional sources of

information, and while the information is obtained in a systematic manner, the scoring may not be standardized.

An *assessment program* or system encompasses more than a single test, instrument, or assessment, usually multiple types of tests and item formats, and is driven by specific goals or a detailed framework. The National Research Council (NRC) (2001) published a report, *Knowing What Students Know: The Science and Design of Educational Assessment,* which advocated for assessment systems to include multiple measures of student performance with a variety of measurement approaches. The Center on Standards and Assessment Implementation (2016) describes that multiple assessments currently include four main types of assessments: formative, diagnostic, interim/benchmark, and summative.

Assessment systems also have to have instructional relevance and should provide information to key stakeholders in the process. For example, the Smarter Balanced Assessment Consortium was funded in 2010 to develop assessments for 3rd thru 12th grade students in English Language Arts and Mathematics. Sireci (2012) describes:

> The assessment system being developed by the Consortium is designed to provide comprehensive information about student achievement that can be used to improve instruction and provide extensive professional development for teachers. The Smarter Balanced assessment system focuses on the need to strongly align curriculum, instruction, and assessment, in a way that provides valuable information to support educational accountability initiatives (p. 4).

Similarly, the NRC (2014) proposes an assessment system for the Next Generation Science Standards that is "composed both of assessments designed to support classroom teaching and learning and those designed for monitoring purposes. In addition, the system should include a

series of indicators to monitor that the students are provided with adequate opportunity to learn science." (p. 193). These two examples highlight the importance of using comprehensive assessment systems to support the teaching and learning of mathematics and science content, while also emphasizing the multi-faceted nature of assessment systems.

Brief Historical Overview

Validity has a nuanced history that at various points in time focused on content, criterion, and construct validities and an evolving future, including an argument-based approaches, new methodologies, and focused attention on the consequences of validity. Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). How to create that evaluative judgment and what evidence is necessary to support the interpretation and use of test scores is still debated among experts.

In a seminal article on construct validity, Cronbach and Meehl's (1955) proposed constructs, which are intangible attributes, are linked with observable attributes. *Educational Measurement* published a series of editions that provide historical context for the development of validity (Cureton, 1951; Messick 1989) and validation (Cronbach, 1971; Kane, 2006). Cureton focused on decisions in selecting the appropriate construct to measure. Cronbach presented test validity as an explanatory empiricism (Markus and Borsboom, 2013) and instead of focusing on what to measure, focused on additional factors that could influence test responses. Messick extended previous work and focused on collecting evidence, arguing for a unified view of validity (rather than the three-prong criterion, content, and construct validities) that includes all evidence needed to use or interpret a test. Then, Kane provided a rich description of a

framework for creating an interpretative argument necessary to justify the use of the test in a specific context. For a detailed evolution on validation see (Messick, 1989; Kane 2006).

Over the last 25 years there has been a diverse array of approaches and frameworks for validation (Kane, 2004; Mislevy, 2006; Pellegrino, DiBello, Goldman, 2016; Schilling, 2004; Wilson, 2005), a focus on validity in educational measurement (Haertel & Lorie, 2004; Kane, 1992, 2001, 2002, 2006, 2013; Mislevy, 1996; Mislevy, Steinberg, & Almond, 2003), frameworks that focus on high-stakes and alternative assessments in education (Haertel, 1999; Marion and Pellegrino, 2006; Oliveri, Lawless, and Young, 2015; Perie and Marion, 2008; Sireci, 2012; Shaw, Crisp, & Johnson, 2012), and principled assessment design approaches (Baxter & Mislevy, 2005; Ferrara, Lai, Reilly, & Nichols, 2017; Wiliam, 2014). Increasingly, for assessments in education, researchers are concerned about consequential validity (Messick, 1989), or using scores in decision making (Mehrens, 1997). This has led to discussions about the use of assessments for teaching and learning, including to inform instruction, aid in decision-making, and to improve learning. Recently, Pellegrino, DiBello, and Goldman (2016) presented a framework that includes instructional validity, which they define as "the extent to which an assessment is aligned with curriculum and instruction, including students' opportunities to learn, as well as how it supports teaching practice by providing valuable and timely instruction-related information" (p. 62). Regardless of the argument used to establish validity, educational researchers should consider the instructional relevance of their assessments, scores, and uses (NRC, 2001).

Five Validation Argument Frameworks

The focus for this section is to provide an overview of five argumentation frameworks for test/instrument validation that have been used within mathematics education scholarship. These

five frameworks will be referenced in the special issue articles. "Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests…[thus], it is incorrect to use the unqualified phrase 'validity of the test'" (AERA, APA, & NCME, 2014, pp. 11-12). To that end, arguments are a "coherent series of reasons, statements or facts intended to support or establish a point of view" (Merriam-Webster, 2018) and provide a rationale for understanding the degree to which something may (or may not) be true. A validation argument functions as an evaluation of the interpretation of test outcomes as well as the actual test scores themselves (Cronbach, 1988; Kane, 2001). The validation argument in general provides instrument users a way to be convinced that the instrument's outcomes actually do what they claim and intend to do. Here, we describe five established validation argument frameworks.

Kane's Interpretation/Use Argument and Validity Argument

Michael Kane is one of the primary proponents in the literature of an argument-based approach to the validation with a strong focus score interpretation and use driving the structure of the argument. Kane's (2001) earlier work discusses an interpretive argument and validity argument, over time Kane's terminology and framework evolve to use the term interpretation/use argument (IUA) and validity argument (VA) (Kane 2012). In 2001, Kane writes that "the main point of the interpretive argument is to make the assumptions and inferences in the interpretation [of test scores/outcomes] as clear as possible" (p. 329). He also described the validity argument as a means to offer a "coherent analysis of the evidence for and against the proposed interpretation and, to the extent possible, the evidence relevant to plausible alternate interpretations" (p. 329). In 2016, Kane provided a slightly different nuanced discussion of IUAs and VAs, which are described below.

Kane's later writings frame an IUA more similarly to the Merriam-Webster definition of argument, "the chain of inferences and supporting assumptions that can get us from the observed test performances to the conclusions of interest and to any decisions to be based on these conclusions" (2016, p. 66). The IUA includes a series of warrants or justification that may be framed as "if-then" statements. There are four inferences: scoring, generalization, extrapolation, and decision (Kane, 2004). The scoring inference connects observed performance and test score. The generalization inference connects an individual's test score to the much larger sample space of possible outcomes on the test from a broader sample. The extrapolation inference connects the test outcome (performance) to an individual's possible outcomes (performance) in the targeted domain of interest. Finally, the decision inference connects an individual's outcome to a decision outcome (e.g., pass/fail). Kane (2006, 2016) recommends an IUA take the form of a Toulmin (1958) model of reasoning with datum, claim, and warrants; in this model, the validation arguments are often presumptive (Kane, 2001) meaning that it is up to test developer to justify their ideas as being appropriate. The VA is "an evaluation of the plausibility of the IUA" (Kane, 2016, p. 69) and presents the evidence that supports the claims (i.e., warrants). Kane's IUA and VA approach has been used to varying degrees by mathematics educators (e.g., Bell et al., 2012; Carney et al., 2017).

In this issue two teams of scholars evaluate the use of Kane's (2004) validity framework, drawing examples from specific instruments in mathematics education. First, Carney, Siebert, Thiede, Crawford, and Osguthorpe discuss the interpretative/use argument from Kane's framework and compare it to an interpretive/use argument based on the sources of evidence from the *Standards* (AERA, APA, & NCME, 2014), described in the next section. Next, Ketterlin-Geller, Perry, and Adams describe aspects of creating an interpretative argument, starting with

inferences and assumptions and ending with intended scores and uses. They integrate the assessment triangle from the NRC (2001) with three of Kane's inferences. A key contribution of their paper is on the instructional decisions that result from test use and on implications for practitioners utilizing such tests.

Sources of Evidence from the Standards Approach

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, 1999) offer a framework for validation arguments through the five sources of validity evidence. While they do not specifically suggest the use of the five sources of validity evidence as a framework for validation, given past approaches to validation it makes sense that test developers would make use of it in this way (e.g., Sireci, 2012). A test does not necessarily need to have evidence for all five sources; the argument should be based on the assumptions/inferences that underlie the proposed score interpretation and use. Five sources are found in both printings: test content, response processes, internal structure, relations to other variables, and consequences of testing. The validation argument is constructed by a test administrator (or user) who aims to justify that the outcomes from the test are sufficiently aligned to the proposed interpretation and use. Test content is framed as "the themes, wording, and format of the items, tasks, or questions on a test" (AERA, APA, & NCME, 2014, p.14). Response processes evidence connects an intended plan for how individuals may respond to an item and the actual ways in which they respond. Internal structure, depending on the approaches taken, has potential to connect the relationships among test items and components to the intended construct, as well as the ways in which score interpretations will be made. Evidence for relations to other variable connects test scores and outcomes to other constructs of interest and has potential to serve as convergent and discriminant evidence. Finally, consequences of testing validity evidence explores the degree to

which anticipated consequences from administering a test (and score outcomes) align with an intended purpose of the test. Mathematics educators (e.g., Bostic & Sondergeld, 2015; Bostic, Sondergeld, Folger, & Kruse, 2017; Gleason, Livers, & Zelkowski, 2017) have used the Standards argument as a means to validate outcomes from two different mathematics tests series (PSMs and MCOP^2). Further, Sireci (2012) utilized the *Standards* framework to provide research recommendations for the Smarter Balanced Assessment Consortium's validation efforts, to ensure the Consortium was meeting its intended goals. As described previously, in this issue Carney et al. compares Kane's approach to an approach that uses the sources of validity evidence from the *Standards* as an approach.

Elemental, Structural, and Ecological Approach

A third approach first described by Schilling and Hill (2007) builds from Kane's (2001; 2004; 2006) notion of IUA and VA. Schilling and Hill initially credit assumptions and inferences with the term *elemental*, "they address the constituent elements upon which the test is based and because validation of this level of assumptions and inference is necessary in order for the rest of the validation process to proceed" (2007, pp. 73-74). The focus at the elemental level is individual test items, themselves. After looking at individual test items through an elemental lens, the next step is a *structural* lens. Structural assumptions and inferences encompass Kane's ideas about generalizability and links between scales and the intended knowledge, traits, or skills that the test claims to measure. The focus at the structural level is the test as a whole. After considering the structural aspects, it necessary to explore a third step: *ecological* aspects.

Since a test has the capacity to impact test takers (e.g., placement into specific courses, further remediation, or identification of an exceptionality) ecological aspects must be explored (Schilling, 2007; Schilling & Hill, 2007). Ecological concerns include aspects of relationships to

other relevant variables as well as some issues of consequences of testing. The elemental-structural-ecological argument framework has been used for the Mathematical Knowledge for Teaching (see Schilling, 2007; Schilling & Hill, 2007; Schilling, Blunk, & Hill, 2007). In this issue, Jacobson and Svetina provide considerations for using the elemental, structural, and ecological approach (Schilling and Hill, 2007), with examples from an instrument designed to measure knowledge and motivation for teaching multidigit arithmetic.

The Assessment Triangle

The National Research Council's report *Knowing What Students Know: The Science and Design of Educational Assessments* (Pellegrino et al, 2001) presented a validation framework, referred to as the assessment triangle, with three interconnected elements: cognition, observation, and interpretation. They claim an assessment cannot be created without consideration of each element. The cognition vertex relates to models of student learning and theories about how students develop understanding of a concept. These models determine what is important to measure in relation to the concept being learned. The observation vertex refers to the description of tasks that will "elicit illuminating responses from students" and is based on assumptions about what is expected students know and are able to do. Finally, the interpretation vertex relates the collected evidence to information about students' understanding and to potential instructional consequences. Taken together, these three elements create an argument about cognitive, instructional, and inferential validity. Recall, in this issue, Ketterlin-Geller, Perry, and Adams integrate the assessment triangle into a discussion regarding an interpretative argument for a formative assessment.

Validity of Instructionally Relevant Assessments

Extending the work of the assessment triangle, Pellegrino, DiBello, and Goldman (2016) detail a validation argument for assessments for classroom use and on the instructional decisions that are made based on the results of the assessment. The framework includes three components: cognitive validity (discussed above), instructional validity, and inferential validity. Instructional validity refers to the degree to which "an assessment is aligned with curriculum and instruction, including students' opportunities to learn, as well as how it supports teaching practice by providing valuable and timely instruction related information" (p. 62). Inferential validity focuses on the "extent to which an assessment reliably and accurately yields model-based information about student performance, especially for diagnostic purposes" (62). This framework provides details on the interpretation of evidence to support teaching and learning. In this issue, Confrey, Toutkoushian, and Shah utilize illustrations from an assessment system to critique a validation argument based largely on Pellegrino et al (2006) framework for the validation of instructionally relevant assessments. They contribute an important discussion on the feasibility of integrating a validation argument into test development.

Contributions of the Special Issue Articles

The manuscripts in this special issue provide an important contribution to the literature through the presentation of argument-based approaches to validation within the context of mathematics education. While argument-based approaches are recommended in the *Standards* (AERA, APA, & NCME, 2014), they are not frequently found in practice (Bostic, Krupa, Carney, & Shih, in press; Cizek et al, 2008). The instances provided in this special issue can serve as a case in point illustration to others, while also providing fodder for discussion around how arguments should be articulated and supported.

One major contribution of this set of manuscripts in the articulation of multiple, example arguments that detail the claims, assumptions, and/or inferences that underlie a test or instrument proposed score interpretation and use. We borrow Kane's term of interpretive/use argument (IUA) to describe this aspect of validation but use it to more broadly to encompass the multiple approaches found in the five different frameworks related to the articulation of the underlying assumptions, inferences, and claims. The IUAs presented in this special issue vary significantly in terms of terminology, structure, and grain size across the manuscripts. While this is to be expected given the lack of clear guidelines and differing frameworks that were used for validation, it raises important questions about the consistency with which validation efforts are implemented. Is there a need for the identification of common elements for IUAs? Is there a way to identify common elements but still allow for flexibility in validation approach? And perhaps most importantly, do we need to give more consideration to the perspective from which an IUA is crafted? For example, in this issue Jacobson and Sevtina's use of Schilling and Hill's (2007) approach highlights the instrument development process, Carney et al.'s use of the *Standard's* (2014) approach highlight the common sources of validity evidence, Confrey et al.'s use of Pellegrino's (2016) approach and Ketterlin-Geller et al.'s use of the NRC's (2001) approach highlight the common categories of inferences/assumptions respectively, and Carney et al.'s use of Kane's (2006, 2013, 2016) approach focuses on the interpretation and use. Each of these perspectives for framing the IUA result in different foci for the articulation of claims, assumptions, and inferences. The field needs to think further about how the use of each of these different perspectives can hide or bring to light important validation considerations, and what that means for recommendations related to validation.

A related aspect to consider is how prescriptive, versus how contingent, validation approaches should be. Kane (2007) articulates this aspect of consideration and is rather clearly in favor of a contingent approach. Jacobson and Svetina (this issue), building upon Schilling's (2004) rationale, explicitly highlight this consideration in their paper, and recommend a more prescriptive approach. Carney's et al. (this issue) close their discussion with a question related to whether the expertise of the instrument developer should be a consideration when determining whether a more prescriptive versus contingent approach should be utilized. Explicit discussion around the affordances and constraints of prescriptive versus contingent approaches, based on IUAs crafted around instruments used in practice, could press both the measurement and mathematics education fields forward in terms of the articulation and use of argument-based validation practices.

References

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

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

Argument. (n.d.). Retrieved from https://www.merriam-webster.com/dictionary/argument

Baxter, G., & Mislevy, R. J. (2005). *The case for an integrated design framework for assessing science inquiry*. Retrieved from SRI International: http://padi.sri.com

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

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

Bostic, J., Krupa, E., Carney, M., & Shih, J. (in press). 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.* New York, NY: Routledge.

Bostic, J. D., & Sondergeld, T. A. (2015). Measuring sixth-grade students' problem solving: Validating an instrument addressing the mathematics common core. *School Science and Mathematics, 115*(6), 281-291.

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.

Carney, M. B., Cavey, L., & Hughes, G. (2017). Assessing Teacher Attentiveness to Student Mathematical Thinking: Validity Claims and Evidence. *The Elementary School Journal*, *118*(2), 281-309.

Center on Standards and Assessment Implementation. (2016). *Overview of major assessment types in standards-based instruction.* San Francisco, CA: WestEd. Retrieved from http://www.csai-online.org/sites/default/files/resources/6257/CSAI_AssessmentTypes.pdf

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68(3), 397-412.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (Vol. 2nd pp. 443-507): American Council on Education.

Cronbach, L. J. (1988). *Five perspectives on validity argument*. In H. Wainer & H Braun (Eds.), Test validity (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 621-694): American Council on Education.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2017). Principled Approaches to Assessment Design, Development, and Implementation. In A. A. Rupp & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 41-74). United Kingdom: John Wiley and Sons.

Gleason, J., Livers, S., & Zelkowski, J. (2017). Mathematics Classroom Observation Protocol for Practices (MCOP2): A validation study. *Investigations in Mathematics Learning, 9*(3), 111-129.

Haertel, E.H. (1999). Validity arguments for high-stakes testing: In search of the evidence. Educational Measurement: Issues and Practice, 18(4), 5–9.

Haertel, E. H., & Lorié, W. A. (2004). Rejoiner to Commentary. *Measurement: Interdisciplinary Research and Perspectives, 2*(2), 129-133.

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

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

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

Kane, M. T. (2006). Valildation. In R. L. Brennan, National Council on Measurement in Education, & American Council on Education (Eds.), *Educational Measurement.* Westport, CT: Praeger Publishers.

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

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17.

Kane, M. T. (2013). The argument-based approach to validation. School Psychology Review, 42(4), 448.

Kane, M. T. (2016). Validation Strategies: Delineating and Validating Proposed Interpretations and Uses of Test Scores. In S. Lane, M. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (Vol. 2nd). New York, NY: Routledge.

Marion, S.F., & Pellegrino, J.W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice*, 25(4), 47–57.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*: Routledge.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: MacMillan.

Mislevy, R. J. (1996), Test theory reconceived. *Journal of Educational Measurement*, 33: 379-416.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. *Educational Measurement, 4*, 257-305.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1).

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing assessments for the next generation science standards*. Washington, DC: The National Academies Press.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research & Perspective, 10*(1-2), 1-29.

Newton, P. E., & Baird, J.-A. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice, 23*(2), 173-177.

Oliveri, M. E., Lawless, R. & Young, J. W. (2015). *A validity framework for the use and development of exported assessment*. Princeton, NJ: Educational Testing Service.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 59-81.

Perie, M., & Marion, S. (2008). Developing a validity argument for a state alternate assessment (AA-AAS) system: A guide for states. Retrieved from http://www.naacpartners.org/projects/valdityGSEG/expertPanel.aspx July 15, 2008.

Schilling, S. G. (2004). Conceptualizing the Validity Argument: An Alternative Approach. *Measurement: Interdisciplinary Research and Perspectives.*

Schilling, S. G. (2007). The role of psychometric modeling in test validation: An application of multidimensional item response theory. *Measurement: Interdisciplinary Research and Perspectives, 5*(2-3), 93-106.

Schilling, S. G., Blunk, M., & Hill, H. C. (2007). Test Validation and the MKT Measures: Generalizations and Conclusions. *Measurement: Interdisciplinary Research and Perspectives, 5(2-3), 118-128.*

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.

Sireci, S. G. (2012). Smarter Balanced Assessment Consortium: Comprehensive Research Agenda. https://portal.smarterbalanced.org/library/en/comprehensive-research-agenda.pdf

Shaw, S., Crisp, V. & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*.

Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in educational and psychological measurement. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 91-111): Springer.

Toulmin, Stephen E. (1958), *The uses of argument*. London: Cambridge University Press.

Wiliam, D. (2014). *Principled assessment design*. London: SSAT.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum.

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice*, 17(2), 117-132.