1-1-2019

Comparison of Two Approaches to Interpretive Use Arguments

Michele Carney
Boise State University

Angela Crawford
Boise State University

Carl Siebert
Boise State University

Rich Osguthorpe
Boise State University

Keith Thiede
Boise State University

Comparison of Two Approaches to Interpretive Use Arguments

Michele Carney*
Boise State University

Angela Crawford
Boise State University

Carl Siebert
Boise State University

Rich Osguthorpe
Boise State University

Keith Thiede
Boise State University

Shortened Version of the Title: Comparison of IUAs

***Name and Address of the Person to Whom Reprint Requests Are to be Sent:**
Michele Carney
Boise State University 1910 University Drive
Boise, Idaho 83725-1745

Keywords: validation, argument, validity, measurement, assessment

Abstract

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) recommend an argument-based approach to validation that involves a clear statement of the intended interpretation and use of test scores, the identification of the underlying assumptions and inferences in that statement—termed the interpretation/use argument, and gathering of evidence to support or refute the assumptions and inferences. We present two approaches to articulating the assumptions and inferences that underlie a score interpretation and use statement, also termed the interpretation/use argument (Kane, 2016). One approach uses the five sources of validity evidence in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) as a framework and the other approach uses Kane's chain of assumptions/inferences approach (Kane, 2006, 2013a, 2016) as a framework. Through this process we identified aspects of these approaches that need to be further clarified for instrument developers to consistently implement either approach, identified important differences in the perspective each approach takes on validation, and highlight important questions for the measurement and mathematics education research fields to consider.

Introduction

Chapter 1 on Validity in the *Standards for Educational and Psychological Testing* (1999; 2014, p. 8) - herein referred to as *Standards* - refers to (a) validity as a quality of the interpretation and use of instrument scores, (b) validation as the process of constructing an argument in support of the proposed interpretation and use, and (c) the argument as explicitly stating and investigating the assumptions/inferences[1] inherent in the proposed interpretation and use. However, not all fields of research apply these conceptions of validity. For example, there is evidence that mathematics education as a research field is not focusing on these conceptions of validity and validation. In a review of articles in

[1] The term assumptions/inferences is used consistently throughout the paper to refer to several terms in the literature, such as assumptions, inferences, claims, and propositions. While there are likely important differences between these terms, they are not used in a consistent manner in the argument-based validation literature. For clarity we use assumptions/inferences throughout but address this inconsistency further in the discussion.

the *Journal for Research on Mathematics Education*, Hill and Shih (2009) found only 8 of 47 studies from 1997 to 2006 provided validity evidence, and that this evidence tended to focus only on internal structure. They found almost no evidence in support of the assumption that instruments actually measured what they intended to measure. Minner, Martinez, and Freeman (2013) investigated the validity and reliability evidence provided for measures used in NSF-funded mathematics and science education research projects from 2008-2012. They describe the validity evidence as having "...an alarming level of missing information (Minner et al., 2013, p. 8)." And perhaps more importantly, the framework Minner et al. (2013) used to develop their synthesis made no mention of validity as a quality of the interpretation and use of instrument scores nor of validation as a process of constructing an argument related to that interpretation and use, indicating these conceptions are not prominent in mathematics and science education research. Lastly, Bostic, Krupa, Carney, and Shih (in press) conducted a review of validity evidence and validation approaches in the *Journal of Research in Mathematics Education* from 1970 to 2017. They found it was relatively common to refer to the validity of an instrument - as opposed to the interpretation and use of scores - and, similar to Hill and Shih's (2009) finding, authors tended to provide a measure or measures of internal structure as the primary supporting evidence for use of an instrument with little focus on validation as an argument. Therefore, while there are examples in mathematics education research that make use of some of the conceptions of validity and validation described in the *Standards* (e.g., Bell et al., 2012; Schilling & Hill, 2007)), these examples are not prevalent.

Issues that may be inhibiting mathematics education researchers from using the conceptions of validity and validation described in the *Standards* are the lack of consensus on how to structure and implement an argument-based approach (Cizek, Rosenberg, & Koons, 2008; Shear & Zumbo, 2014) and/or the lack of examples situated in mathematics education that make use of an argument-based approach (Bostic et al., in press). One way to ensure an argument-based validity approach encompasses the breadth and depth of the assumptions/inferences that underlie an interpretation and use statement, is to have well- delineated processes for the articulation of (and resulting structure for) the assumptions/inferences inherent in a proposed interpretation and use. The purpose of this paper is to present examples for the articulation of assumptions/inferences for a validation argument using two different approaches. In presenting these examples, our goals are to: (a) provide case illustrations of the articulation of assumptions/inferences to assist test developers in specifying their own IUAs, (b) examine the resulting articulation of assumptions/inferences to better understand if certain assumptions/inferences are highlighted or hidden by a particular validation framework, and perhaps most important, (c) provide a framework for discussion in the instrument development and mathematics education communities around how to structure these types of arguments and further make use of current conceptions of validity and validation. In what follows, we first provide an overview of argument-based validation approaches. Second, we generate an interpretation/use argument for two different approaches to validation and describe the process of generating that argument. Finally, we analyze these processes of generating arguments and discuss considerations for improving their structure in future research.

Overview of Argument-Based Validation Approaches

While discussion is ongoing in the measurement community around concepts of validity and validation (for an example see volume 23, issue 2 of *Assessment in Education: Principles, Policy, and Practice*), we are starting from the shared vision expressed in the *Standards* (2014). As stated previously, this shared vision focuses on: validity as a quality of the interpretation and use of instrument scores; validation as the process of constructing an argument in support of the proposed interpretation and use; and the argument as an explicit statement and investigation of the assumptions/inferences inherent in the proposed interpretation and use. This section provides theoretical background for the conceptualization of validation, situates three argument-based perspectives on validation, describes the importance of the articulation of assumptions/inferences that underlie an interpretation/use argument, and introduces an example instrument on which to base the case illustrations to be examined in this paper.

Theoretical Background

Historically, validation has often been conceptualized in practice through a 'category' approach which involves presenting evidence from one or more genres of validity. Using this conceptualization, it was commonplace for researchers and practitioners to present a category or categories of validity evidence (e.g., content, criterion, or construct validity, etc.) and describe this presentation of evidence as connoting 'test validity' (Cizek et al., 2008). The concern with this approach was that it often situated the test as having validity and it did not focus on justifying score interpretation(s) for a particular use or uses (Cronbach, 1988; Kane, 1992; Messick, 1989). Bachman (2005) describes this as an ad hoc approach in which "...test developers in practice have typically adopted a weak program of construct validation that may consist of whatever evidence is relatively easy to collect that provides support for the intended

interpretation (Bachman, 2005, p. 30)." Messick's (1989) unified model of validation was intended to address these issues by providing a more comprehensive approach; however, the open-endedness of the process made implementation difficult. Kane (1992, 2013a) describes argument-based processes as providing a more practical process for validation. However, there is evidence that practice continues to lag behind theory in terms of argument-based approaches to validation (Cizek et al., 2008; Shear & Zumbo, 2014; Wolming & Wikström, 2010).[2]

Three Perspectives

Argument-based approaches to validation can be situated in three perspectives: validation via principled design, common identified categories, and chain of assumption/inferences. Validation via principled design is a broad category that encompasses several approaches to assessment design and validation that share common foundational elements (see Ferrara, Lai, Reilly, & Nichols, 2016 for a description of these elements and related approaches). Validation via principled design is particularly relevant for assessments designed to make examinee cognition explicit. In terms of validation, the argument development and evidence collection is built into the assessment development process.

Approaches that focus on common identified categories involve using a framework of pre-identified categories of evidence or assumptions/inferences that have been generalized as important for particular types of assessments (e.g., Pellegrino, DiBello, & Goldman, 2016 for instructionally relevant assessments) or across all types of assessments (e.g., Schilling & Hill, 2007), and then identifying the assumptions/inferences specific to the particular test interpretation and use within each category. Schilling and Hill (2007) in their articulation of assumptions/inferences for the Learning Mathematics for Teaching instruments used this approach when they identified three general categories of assumptions: elementary, structural, and ecological. Next, they identified assumptions and inferences within each of these categories specific to their instrument. Similarly, Bell et al. (2012) focused on presenting evidence within the categories of inferences described by Kane (2006): scoring, generalization, extrapolation, and implication. More recently Pellegrino et al. (2016) identified three components of validity: cognitive, instructional, and inferential specific to validation involving instructionally relevant assessments. A potential advantage of the category approach is that it provides an explicit structure for thinking about the articulation of assumptions/inferences. Kane (2007) expressed concern related to the common identified categories approach because he views it as too prescriptive in that it '...will not apply equally well to all cases and may be applied without fully considering their appropriateness for a particular use, in a particular context, with a particular population (Kane, 2007, p. 182).'

Kane (2007) describes a more contingent approach that involves clearly stating the interpretation and use for an instrument and then laying out a chain of reasoning that articulates assumptions/inferences starting at the test response and leading all the way to score interpretation and use. This chain of reasoning and associated assumptions/inferences are termed the interpretation/use argument (IUA). Kane (2001, 2004, 2006) suggests the IUA is used to identify and prioritize the most critical aspects of evidence that must be collected in support of the IUA. The evidence then collected to examine the assumptions/inferences in the IUA is termed the validity argument. Kane's perspective situates validation in the context of the specific instrument. He recognizes a more contingent approach also has its potential downfalls in that it '…is open to variation across test developers and validators in how they might decide to state and validate a proposed interpretation or use (Kane, 2007, p. 182)' but feels the contingent approach is more appropriate in terms of situating the argument in the context of the score interpretation and use specific to the instrument being examined.

The Importance of Articulating Assumptions/Inferences

The articulation of assumptions/inferences that underlie a proposed score interpretation for a particular use is important; it serves as a key initial element to the development of a comprehensive argument. However, there is extensive variability in the approach to developing, and the resulting structure of, IUAs. From Kane's perspective, the proposed interpretations and uses for instruments vary widely, and therefore the underlying assumptions/inferences that need to be identified also vary widely. However, others (e.g., Ferrara et al., 2016; Schilling & Hill, 2007) feel a more prescriptive approach is needed to ensure important aspects of validation are not overlooked. An important

[2] It is important to note that this summation of the 'category' approach is based on what has often been published in practice and it may not represent what was intended by theory. Therefore, it is important to determine whether argument-based approaches implemented in practice are a closer match to intended theory. In other words, is it the theory that needs to change or is it actually the implementation of validation approaches in practice where issues arise?

methodological question for an argument-based validation approach to address is how to ensure the appropriate and critical assumptions/inferences have been identified for a particular instrument. We are extending the use of the phrase IUA to encompass both the more prescriptive category approach and the more contingent chain of reasoning approach to the articulation of assumptions/inferences.

The *Standards* provide an example of a set of identified assumptions/inferences (they use the term propositions):

> 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 for the particular purpose of testing. For instance, when a mathematics achievement test is used to assess readiness for an advanced course, evidence for the following propositions might be relevant: (a) that certain skills are prerequisite for the advanced course; (b) that the content domain of the test is consistent with these prerequisite skills; (c) that test scores can be generalized across relevant test sets of items; (d) that test scores are not unduly influenced by ancillary variables, such as writing ability; (e) that success in the advanced course can be validly assessed; and (f) that test takers with high scores on the test will be more successful in the advanced course than test takers with low scores on the test (p. 12).

Yet the process of identifying the assumptions/inferences, and more importantly, how to ensure all appropriate assumptions/inferences have been identified, is not clear.

The purpose for this paper is to present IUAs for the same instrument using two different approaches[3]. One IUA will be developed and framed using a common identified category approach with the sources of validity evidence described in the *Standards* (2014) as the categories (see Sireci, 2012 as an example of the Standards used as a framework for validation). While the *Standards* do not explicitly advocate this as a validation approach, we see it as a likely interpretation given past validation approaches focused on presenting evidence from particular categories of validity as connoting validity for an assessment and recommendations from theorists related to this idea (e.g., Ferrara et al., 2016). The second IUA will be developed and framed through Kane's (2004; 2006, 2013a, 2016) chain of assumptions/inferences approach starting at the item response and ending as score interpretation for a proposed use.

Example Instrument: The Diagnostic Assessments of Proportional Reasoning

To develop the two example IUAs using the category and chain of reasoning approaches, we need an example instrument. The Diagnostic Assessments of Proportional Reasoning (DAPR) serves as the example to frame our two IUAs. We selected the DAPR due to the experience of two of the authors in developing the assessments and interest in investigating the validation of the interpretation and use of the assessments. The structure of an IUA is highly dependent upon the particular assessment's interpretation and use. Therefore, selecting a different instrument would likely results in very different IUAs. Our intent is to provide a case in point example, with the understanding that our findings related to the IUA structure are tightly connected to the stated interpretation and use of the instrument. While ideally the IUA is articulated prior to initial assessment development, in our case it is being used to iteratively examine and improve assessments that have been previously developed.

The DAPR are designed to assist teachers in diagnosing student understanding related to composed unit and multiplicative comparison conceptions (Lobato, Ellis, & Charles, 2010) in emergent proportional reasoning situations. The DAPR assessment content is targeted at middle grades standards in the Common Core State Standards for Mathematics (NGA & CCSSO, 2011) and has been vetted with students in grades 6-9. The DAPR is a 20 item fill-in-the-blank assessment available in three equated forms. Hardcopy forms are administered by classroom teachers with a 20 minute time limit. The forms are scored as the sum of the number correct and results are interpreted and used by the classroom teacher.

> The DAPR Interpretation and Use Statement asserts:
>
> A student's DAPR score can be interpreted in relation to a hypothetical trajectory of composed unit and multiplicative comparison understanding (M. Carney & E. Smith, 2017; M. B. Carney & E. Smith, 2017). The scores can be used by classroom teachers to identify instructional needs for

[3] The validity by design approach dictates the development and validation process must be done in conjunction with one another. The example assessments were not developed via this framework; therefore, this approach would not apply well. In addition, there are multiple examples of validity by design being used in practice so there is less need for examples illustrating their IUAs.

> students (Use 1) and could be used as one of multiple measures to identify students in need of remediation (Use 2). The scores should not be used in isolation to identify students in need of remediation.

The above theoretical background on validity, perspectives on argument-based approaches, stated importance of articulating assumptions/inferences, and description of an example instrument provide a framework from which to generate the IUAs that follow.

Generating an IUA Using Kane's Chain of Assumptions/Inferences Approach

Michael Kane has written extensively about argument-based approaches to test validation (see for example Kane, 2001; Kane, 2004; Kane, 2006, 2013b, 2016). His writings describe a validation approach in which a chain of reasoning is articulated involving the statement of assumptions/inferences which links a person response to a score interpretation for a specified use. Kane's perspective situates validation in the context of the specific instrument. He sees the chain of reasoning involving the articulated assumptions/inferences as highly dependent upon the intended score interpretation and use and rejects the notion of a one-size fits all approach to validation as suggested by a more prescriptive (i.e., categories) approaches:

> ...this set of inferences [scoring, generalization, extrapolation, theory-based inferences, score use] is not intended as a checklist...The IUA is to specify the proposed interpretation and use as it is to apply to the populations and in contexts in which it will be used, and is to do so in enough detail to provide a framework for the evaluation of its most critical and questionable inferences and assumptions. It does not need to follow any particular pattern (Kane, 2016, p. 71).

Instrument developers have used Kane's approach or aspects of Kane's approach to present validation arguments for instruments. Some developers describe having focused on aspects of Kane's approach but do not explicitly articulate the assumptions/inferences that underlie the score interpretation and use from the perspective of a linked chain of reasoning from person response to score interpretation and use. For example, Ruiz-Primo et al. (2012) in their interpretative argument for instructionally sensitive assessments, articulate two important assumptions associated with their score interpretation and use. However, these assumptions are not articulated from the perspective of developing a chain of reasoning that starts at the student response leading to the interpretation and use. Others have opted to use the common categories of assumptions/inferences described by Kane (scoring, generalization, extrapolation, decision) as the focus of their framework for validation (Bell et al., 2012; Clauser, Margolis, Holtman, Katsufrakis, & Hawkins, 2012). While it may be that this generic framework of assumptions/inferences serves to identify the critical assumptions/inferences that underlie a score interpretation and use, it does not place in the forefront what appears to be the crux of Kane's approach - the need to situate the argument with the specific context of the score interpretation and use. And while there are examples that seem to adopt the crux of Kane's approach, embedding the articulating of the assumptions/inferences for the IUA as a chain of reasoning in the specific context of the score interpretation and use (e.g., Chapelle, Enright, & Jamieson, 2010), these are not commonly found in the literature, particularly in the context of mathematics education. Thus, the need exists to provide an example of an IUA, situated within the context of mathematics education that makes use of Kane's chain of assumptions/inferences approach.

To generate our IUA using Kane's chain of assumptions/inferences approach, we engaged in the following analysis. First, we developed our interpretation and use statement for the DAPR assessments (see section on Diagnostic Assessments of Proportional Reasoning). Our focus was on identifying the assumptions/inferences that underlie the DAPR Interpretation and Use Statement - starting at student generated responses to assessment items and then linking the assumptions/inferences all the way to the specific interpretation and then use of the assessment scores that were stated. We used Kane's common inference categories to assist in framing this work, but they were a secondary to ensuring our assumptions/inferences linked appropriately to provide a sufficient chain of reasoning, first to the proposed interpretation, and then to the two stated uses. We organized the IUA with the articulated chain of assumptions/inferences in Table 1.

[Insert Table 1 Here]

Through the process of generating an IUA using Kane's chain of assumption/inferences approach, we identified several points of discussion. While there is not room to discuss all the points, we highlight two important points specific to using Kane's approach--structure of the assumption/inference statements and terminology usage, and starting point of the argument.

Structure of Assumptions/Inferences Statements & Terminology

An important point of discussion that arose while using Kane's chain of assumptions/inferences to articulate the IUA was the structure of the statements. Focusing on identifying assumptions/inferences inherent in going from the observed performance to a score (we actually backed up a step and started at test administration leading to observed performance) provided a helpful starting place. However, what needed to be included in the assumption/inference statement and the granularity with which it should be framed became very muddled as we looked at examples (Bell et al., 2012; Chapelle et al., 2010; Kane, 2004) and discussions in the literature (Kane, 2007; Schilling & Hill, 2007) for guidance. Essentially inconsistencies in how these statements are constructed and varying levels of granularity made it difficult to understand how best to structure them. In addition, as previously stated, the language usage in the literature is relatively inconsistent in regards to what is being discussed, particularly in relation to the terms claims, propositions, assumptions, and inferences. For example, the interpretative argument presented by (Bell et al., 2012) is relatively concise - four generic inference categories from Kane (2006) with 10 related statements - and the terminology primarily refers to inferences with the occasional inclusion of the term assumptions used in an interchangeable manner. Other the other hand, (Chapelle et al., 2010) provide a rather detailed argument - six inference categories each with a stated warrant licensing the inference and 17 assumption statements underlying each inference - and multiple terms are used - inference, assumption, warrant, and claim - each in relation to specific aspects of the argument.

We structured our statements in the format of the first sentence expressing the assumption being made when going from one link in the chain to the next (e.g., observed performance to scoring) and the next sentence expressing the inference that could be made based on that assumption, with the assumption serving as the aspect of the statement that could be supported with empirical evidence. While we agree with Kane's position that the structure of the argument should be flexible based on the context in which the instrument will be interpreted and used, we feel that clear guidelines related to structure of the statements could be provided to instrument developers and that this would not hinder the flexibility that Kane is looking for in how arguments are developed.

Starting Point of the Argument

While the recommendation to start the first link of the argument from the observed performance (or in our case the test administration) was helpful in terms of providing a clear starting point, it felt disconnected from what is typically the focus of the initial development process – clear articulation of the construct to be measured in conjunction with how to meaningfully operationalize that construct into items and an instrument. This disconnect caused two primary issues for us in the development of our IUA. First, it did not put high-quality operationalization of the construct at the forefront of the validation process. Given that ensuring we had operationalized the construct in a meaningful way had been a multi-year process, the lack of a clear focus on this aspect of development within the argument gave the impression that a critical aspect of validation was missing. Second, it separates test development from validation in a way that made validation feel like a separate process. This perceived disconnect may have been the result of constructing the IUA following the assessment development process. Kane indicates the IUA should be constructed prior to development (Kane, 2016) so this may have been an issue with our ordering of the assessment and IUA development. However, we see it as more likely that construct representation should fit within Kane's suggested inferences of extrapolation to theory-based interpretation (the theory-based interpretation inference appears in later writings, e.g., Kane, 2013a). Given the ordering of the test development process this feels at odds with how instruments are constructed. Wilson (2018) addressed this disconnect by recommending a test development argument occur separately or prior to the interpretive use argument. While Shaw et al (2012) suggested adding an inference at the beginning of the argument specific to construct representation. However, it may be this disconnect is a result of our tendency to prioritize test development over the test interpretation/use. With the shift in validity from being a quality of the test to a quality of the score interpretation and use, it may be that Kane's approach is intentional in its lack of strong focus on construct representation in order to prioritize interpretation and use.

Generating an IUA Using the *Standards* Sources of Validity Evidence Approach

The *Standards* provide guidance to test developers and users on issues related to validity and validation. They call for developers to clearly state the intended interpretation(s) and use(s) for an instrument, and describe validation as "...a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use (p. 11)." In addition, they call for the development of "... a set of propositions or claims that support the proposed interpretation for the particular purpose of testing (p. 12)." They provide some example assumptions/inferences (they use the terms propositions or claims) for assessments and suggest considering rival hypotheses with a focus on potential issues of construct underrepresentation and construct irrelevance to assist in the generation of assumptions/inferences. We see this concept of generating assumptions/inferences as similar to what Kane (2013, 2016) describes as an IUA, exclusive of the focus on developing a comprehensive argument via a chain of linked assumptions/inferences. Once the IUA has been generated, the *Standards* indicate validation can proceed in terms of obtaining empirical evidence, again similar to what Kane (2013; 2016) describes at the validity argument.

We see it as likely that instrument developers could view the five sources of validity evidence from the *Standards* as a framework in which to think about the evidence they need to provide in support the of intended interpretation of test scores for a proposed use. While this is not explicitly suggested by the *Standards*, given the historical tendency and even current recommendations to provide validity evidence within certain categories (e.g., DeVellis, 201?), we see it as likely. For example, (Sireci, 2012) used the sources as the validation framework for the presentation of evidence for the Smarter Balanced Assessment system. Given the prominence of the *Standards* and because we find it likely instrument developers might use this approach, providing an IUA developed through the sources of validity evidence as a framework should prove useful both as a discussion point in the literature but also as a case illustration to those interested in conducting similar work.

To generate our IUA using the sources of evidence from the *Standards* as a framework we engaged in the following steps. First, we developed our interpretation and use statement for the DAPR assessments (see section on Diagnostic Assessments of Proportional Reasoning). We then used an iterative process of carefully reading the sources of validity evidence in the *Standards*, in conjunction with the DAPR Score Interpretation for Proposed Uses statement, to construct assumption statements within each source of validity evidence. This was the bulk of the work involved in articulating the IUA. Following an initial discussion of assumptions/inferences within each source of validity evidence, we then followed the recommendations from the *Standards* by focusing on generating rival hypotheses with a focus on construct irrelevance and underrepresentation to try to identify additional assumptions/inferences within the sources of evidence categories. Lastly, we organized the identified assumptions/inferences in a table framed around the sources of evidence. Table X provides the resulting IUA using the sources of evidence from the *Standards* as a framework. Through this process, we found that each source required the articulation of multiple assumptions/inferences. In addition, we found that particular assumptions/inferences stood out to us as more critical in relation to the proposed interpretation and use. Therefore, Table 2 presents the selected assumptions/inferences we deemed most critical to support the DAPR Score Interpretation for Proposed Uses statement.

[Insert Table 2 Here]

Through the process of generating an IUA using the sources of evidence from the *Standards* as a framework, we identified several points of discussion. While there is not room to discuss all the points, we highlight the two points specific to using the sources of evidence in the *Standards* as a framework: (a) 'fit' within a category, and (b) similar to Kane's approach, quality and depth of construct representation.

'Fitting' Assumptions/Inferences within a Category

One point of discussion that arose both during the initial generation of assumptions/inferences and during the discussion related to rival hypothesis was where particular assumptions/inferences 'fit' within the sources. For example, the assumption *'The item type framework and associated items adequately measure composed unit and multiplicative comparison conceptions'* fits within the test content category based on the idea that evidence needed to support this assumption would demonstrate a relationship between the content of a test and the construct it is intended to measure. However, based on previous experience in attempting to assess composed unit and multiplicative comparison conceptions, we knew we would need to generate evidence from individual student cognitive interviews (response process evidence) to demonstrate the relationship between the items types and the conceptions we intended to measure. Therefore, this particular assumption fit within two categories of evidence. This was common across

several of the assumptions/inferences where - once generated based on thinking about a particular evidence source - we had difficulty agreeing upon where they fit in the evidence source framework. While this is not necessarily problematic, because the end goal is to generate critical assumptions/inferences, the lack of clean fit within a particular source of evidence seems like an important point of discussion to consider in terms of how the argument is structured and presented to others.

Quality and Depth of Construct Representation

Another point of discussion that arose was related to an assumption we were aware was particularly critical prior to beginning the process of articulating IUAs through either of the frameworks. The assumption was *'Responses to items reflect proportional reasoning understanding of composed unit and multiplicative comparison conceptions and do not unduly reflect construct irrelevant strategies, such as input-output (correspondence based) strategies or more general test-taking strategies.'* Based on previous attempts to develop items that assessed students' composed unit and multiplicative comparison understanding (Carney, Smith, Hughes, Brendefur, & Crawford, 2016), we knew that assessing these conceptions was particularly difficult and that strong evidence would be needed to support this statement. While this clearly fits within the response process source of evidence, we wondered if test developers and/or users without a strong background on student cognition related to proportional reasoning would necessarily have recognized the need to explicate this interpretation and related assumption so explicitly. It highlighted our perception that we tended to frame sources of evidence related to content around two ideas: (a) constructs that are easily defined and mapped, and (b) that which a developer would typically need in order to justify how they sampled the domain (if the domain was relatively well defined). This is in comparison to cognitive constructs which at times are ill- defined and for which there needs to be strong theory and related evidence regarding how the construct is defined and assessed. For the latter situations, the evidence category of content did not connote the amount of type of evidence necessary to justify that an 'ill-defined' cognitive construct has been appropriate and thorough defined and represented.

Discussion

Kane (2007) used the terms contingent and prescriptive to describe two ends of a continuum related to validation approaches. The common identified categories approach falls on the prescriptive end of the continuum because articulation of the IUA is focused on addressing pre-specified categories (e.g., the five Sources of Validity Evidence from the S*tandards*). The use of a linked chain of assumption/inferences from response process to interpretation and use falls on the contingent end of the continuum because articulation of the IUA is highly contingent upon the stated interpretation and use for the particular instrument. Throughout the development of the two IUAs a consistent theme was where an approach, or aspect of an approach, fell along the continuum from contingent to prescriptive. In the first part of the discussion we provide four recommendations for clarifying the structure of IUAs or aspects of the approaches. The second part of the discussion compares the perspectives from which the two approaches approach validation efforts. The final section addresses a broader perspective related whether or not recommendations related to prescriptive versus contingent approaches should take into the expertise of the developer.

IUA Structure Clarifications

The following bulleted list of recommendations address critical aspects of the IUA development process that are in need of clarification from those in the measurement and/or mathematics education community. Clarifying these aspects would assist instrument developers in implementing appropriate interpretation and use of argument-based approaches. It is important to note that while these recommendations for clarification make aspects of the approaches more structured, they do not reduce the contingent nature Kane prioritizes. There is a need to:

A. Clearly state the interpretation and use as the first step in the IUA development process. While this is highlighted in the Standards (2014) and Kane (2004, 2006, 2012, 2013, 2016), it is not clearly and consistently done in the literature.

B. Consistently use terminology related to the articulation of an IUA. Currently practice involves inconsistent use of terms such as assumptions, inferences, claims, warrants, and propositions. These terms need to be clearly defined and consistently used in the articulation of IUAs.

C. Clarify where operationalization of the construct fits into the IUA structure (e.g., Shaw, Crisp, & Johnson, 2012). As Chapelle (2012) recommends "What needs to be better understood is how the role of the construct in the interpretive argument can and should influence the way it is defined (p.24)."

D. Provide guidance in terms of the structure and level of granularity for the articulation of assumptions/inferences.

The four clarifications above are necessary in order for developers to craft IUAs with some level of consistency across instruments. Consistency is needed to ensure developers craft IUAs that identify critical assumptions/inferences and can be understood by others. If we do not move towards clarifying these aspects of argument-based approaches to validation, we run the risk of having IUAs crafted with widely varying levels of specification and focus, which could lead to instruments whose score interpretation and use is not well specified or supported.

Comparison of the Approaches

In this case in point example, the contingent approach highlighted more assumptions related to the interpretation and use in a particular context and placed less emphasis on the development of the instrument itself. In contrast, the more prescriptive approach highlighted aspects of the validation process that are more commonly focused on in research, such as operationalizing the construct, and content and structural features of the instrument and placed less emphasis on practical implementation. While the process of collecting evidence for arguments from either of these two approaches may reinforce these differences or bridge the "gap", this initial step of articulating an argument seemed to result in a different focus which somewhat changes the perspective on validity. We feel this difference in focus is an important one that merits further exploration.

Expertise of the Developer

Returning to the theme of prescriptive versus contingent approaches, the process of articulating the two example IUAs left us wondering if the recommendation to use a particular validation approach should in some way be dependent upon the expertise of the instrument developer. We posit that a significant number of instrument developers do not necessarily have a high level of expertise in validation, and in particular more modern conceptions of validation.

For example, consider a psychometrician who developers assessments for an international large testing company, a faculty member who is a disciplinary expert (e.g., mathematics education) and who may be designing assessments to assess the outcomes of a grant-funded mathematics intervention, or a school district math coordinator who is charged with creating math placement tests that will be used to determine what math track (on, above, or below grade-level) student are on in middle school. These situations are high stakes in terms of funding and long-term student outcomes, respectively. However the level of expertise these individuals would have in terms of validation is likely to be varied. For example, while a psychometrician may have more expertise in modern approaches to development and validation, a school district math coordinator may have more focus on and understanding of how the instrument will be interpreted and used. As we think about the potential affordances and constraints of contingent versus prescriptive approaches - do we need to take into account the expertise of the test developer and validator when thinking about recommendations for argument-based validation approaches? We do not pretend to know the answer to this question but posit to the measurement and mathematics education that the expertise of the test developer and validator may be an important factor to consider when developing and/or describing contingent versus prescriptive approaches to validation.

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*: American Educational Research Association.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*: American Educational Research Association.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal, 2*(1), 1-34.

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. D., Krupa, E., Carney, M., & Shih, J. (in press). Reflecting on the past and thinking ahead in the measurement of students' outcomes. In J. D. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge*. New York, NY: Routledge.

Carney, M., Smith, E., Hughes, G., Brendefur, J., & Crawford, A. (2016). Influence of Proportional Number Relationships on Item Accessibility and Students Strategies. *Mathematics Education Research Journal*. doi:10.1007/s13394-016-0177-z

Carney, M., & Smith, E. (2017). *Analyzing Item Measure Hierarchies to Develop a Model of Students' Proportional Reasoning*. Paper presented at the American Education Research Association Annual Meeting, San Antonio, TX.

Carney, M. B., & Smith, E. (2017). *Using Instrument Development Processes to Iteratively Improve Construct Maps: An Example in Proportional Reasoning*. Paper presented at the NCME Special Conference - Classroom Assessment and Large-Scale Psychometrics: The Twain Shall Meet, Lawrence, Kansas.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing, 29*(1), 19-27.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

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.

Clauser, B. E., Margolis, M. J., Holtman, M. C., Katsufrakis, P. J., & Hawkins, R. E. (2012). Validity considerations in the assessment of professionalism. *Advances in health sciences education, 17*(2), 165-181.

Cronbach, L. J. (1988). Five perspectives on validity argument. *Test validity*, 3-17.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2016). Principled Approaches to Assessment Design, Development, and Implementation. *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*, 41.

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. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 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. doi:10.1080/15366360701492807

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

Kane, M. T. (2013b). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement, 50*(1), 115-122. doi:10.1111/jedm.12007

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.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Minner, D., Martinez, A., & Freeman, B. (2013). *Compendium of Research Instruments for STEM Education, Part 1: Teacher Practices, PCK and Content Knowledge with Addendum*. Retrieved from http://cadrek12.org/resources/compendium-research- instruments-stem-education-part-i-teacher-practices-pck-and-content-kn

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.

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. doi:10.1080/15366360701486965

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, 19*(2), 159-176.

Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in educational and psychological measurement *Validity and validation in social, behavioral, and health sciences* (pp. 91-111): Springer.

Sireci, S. G. (2012). *Smarter balanced assessment consortium: Comprehensive research agenda*. Retrieved from https://portal.smarterbalanced.org/library/en/comprehensive-research- agenda.pdf

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

Table 1. IUA Using Kane's Chain of Assumption/Inferences Approach.

Kane's Common Inferences	Assumption/Inference Statements
Administration to Observed Performance	The rules of administration are clearly articulated, and administration is appropriately implemented by classroom teachers. Therefore, the observed performance was not unduly influenced by setting or other situational factors.
Observed Performance to Scoring	Classroom teachers can apply the articulated DAPR scoring rules to the observed performance. Therefore, a student's DAPR score is an accurate reflection [or operationalization] of the observed performance.
Scoring to Generalization	A student's DAPR score is an accurate estimate of student performance across similar item types, contexts, and number relationships. Therefore, a student's DAPR score is relatively equivalent across test forms.
Generalization to Extrapolation	The item types, contexts, and number relationships that make up the test domain appropriately capture emergent proportional reasoning related to students' composed unit and multiplicative comparison conceptions. Therefore, a student's DAPR score is reflective of emergent composed unit and multiplicative comparison conceptions.
Extrapolation to Theory-Based Interpretation	The composed unit and multiplicative comparison conceptions are hierarchically ordered. Therefore, a student's DAPR score can be interpreted in relation to a continuum of students' composed unit and multiplicative comparison conceptions.
Theory-Based Interpretation to Use 1	A student's instructional needs related to emergent proportional reasoning are appropriately identified. Therefore**,** a teacher can use the student's DAPR score to identify their instructional needs.
Generalization to Relations	A student's composed unit and multiplicative comparison conceptions are related to their (a) general proportional reasoning ability, and (b) general mathematics achievement. Therefore, the DAPR assessment is predictive of important aspects of mathematics achievement.
Relations to Use 2	DAPR scores are predictive of general mathematics achievement. Therefore, a student's DARP score can be used as one of multiple measures to identify students in need of remediation.

Table 2. IUA Using the Sources of Validity Evidence from the *Standards* to Frame the Articulation of Assumptions/Inferences.

Source of Validity Evidence	Selected Assumptions/Inferences Statements
Test content: relationship between the content of a test and the construct it is intended to measure	-The item type framework and associated items adequately measure composed unit and multiplicative comparison conceptions. -Other factors known to influence item difficulty (e.g., item format, number sets, contexts used, number relationships) were controlled for in a way that allows for the isolation of composed unit and multiplicative comparison conceptions. -The total correct can be appropriately interpreted along a continuum of item types and student conceptions.
Response Processes: Cognitive processes engaged in by test takers	-Responses to items reflect proportional reasoning understanding of composed unit and multiplicative comparison conceptions and do not unduly reflect construct irrelevant strategies, such as input-output (correspondence based) strategies or more general test-taking strategies. -The order of the test items is not unduly influencing the response process.
Internal Structure: Test items and structures conforming to construct	-Composed unit and multiplicative comparison understanding are unidimensional. -The learning trajectory for students' composed unit and multiplicative comparison conceptions is hierarchically structured with composed unit conceptions typically occurring prior to multiplicative comparison conceptions. -The three DARP forms are equivalent in difficulty.
Relations to Other Variables	The scores on the DAPR test are: ➔ Predictive of general proportional reasoning ability and overall mathematics achievement ➔ Less highly related to the ability to algebraically manipulate equations than to general proportional reasoning
Consequences of Testing: Direct results from the interpretation of test scores	-The score interpretation provides a mechanism to appropriately differentiate instruction -Teachers will not unduly narrow instructional focus to items constructed from the item type framework