Assessing Measures of Mathematical Knowledge for Teaching: A Validity Argument Approach

Stephen G. Schilling \(^a\); Heather C. Hill \(^b\)

\(^a\) University of Michigan, Ann Arbor, Michigan, USA
\(^b\) Harvard Graduate School of Education, Cambridge, Massachusetts, USA


URL: http://dx.doi.org/10.1080/15366360701486965
Assessing Measures of Mathematical Knowledge for Teaching: A Validity Argument Approach

Stephen G. Schilling
University of Michigan, Ann Arbor, Michigan

Heather C. Hill
Harvard Graduate School of Education, Cambridge, Massachusetts

In assessing the utility of a test, two issues stand out: whether it provides information of interest to test consumers, and whether scores generated by the test assist in making good decisions. Validity addresses these two issues, making an assessment of test validity the single most important product provided by test developers. Unfortunately, despite its importance, test validation is almost universally viewed as the most unsatisfactory aspect of test development. As Messick (1988) noted, there has been a consistent disjunction between validity conceptualization and validation practice. To start, the proliferation of many different kinds of validity evidence without clear prioritization presents test consumers with an enormous task, that of sifting through various methods, approaches, and empirical work to determine the usability of a test. At the same time, some test developers use evidence (and methods) selectively, choosing convenient means for test validation, and convenient results for reporting.

Kane (2001, 2004a) developed an argument-based approach to validity as a means of addressing these difficulties. His approach consists of two stages:

- The Formative Stage: Developing an interpretive argument, in which the assumptions and inferences involved in the proposed interpretation of the test are stated explicitly.

Correspondence should be addressed to Stephen G. Schilling, School of Education, University of Michigan, 1160 SEB, Ann Arbor, MI 48128. E-mail: schillsg@umich.edu
• The Summative Stage: Developing the validity argument, in which the interpretive argument is evaluated and possibly reformulated in light of assembled empirical evidence.

In this framework, the interpretive argument provides both the basis for organizing and prioritizing validity evidence and a means for gauging the progress of the validation effort.

While Kane’s validity argument approach is widely viewed as a positive development, there have been few examples of its implementation. In the absence of real-world examples, Kane’s approach becomes yet another case of the phenomena that concerned Messick; the disjunction between validity conceptualization and practice. Further, without a real-world implementation, we cannot assess the efficacy of the Kane approach, or answer questions that have arisen about the structure of such an argument.

In this set of papers, we use Kane’s approach to validate a measure of teachers’ mathematical knowledge. We do so not only to learn about the measure itself, but also to assess the promise of and problems with the argument-based approach to validity. In doing so, we attend closely to issues raised in the lead author’s and others’ responses to Kane’s original work.

One such issue is whether interpretive and validity arguments are necessarily unique to every test, as Kane asserts, or whether there is a general structure that must guide any approach to test validation. Schilling (2004) suggested, in his response to Kane, that a better approach would be more prescriptive, describing important types of assumptions and inferences common to all tests and structuring validity arguments around these types of assumptions and inferences. In this set of papers, we propose three such types of validation-related assumptions and inferences: elemental, or assumptions and inferences concerning the performance of specific test items, including consistency of items with subjects’ knowledge; structural, or assumptions and inferences concerning the internal structure of the test, including the consistency of the structure of the test with the structure of the test domain; and ecological, or assumptions and inferences concerning the external structure of the test, including the relationship of the test scales with external variables. Our work validating the mathematical knowledge measures suggests that absent such a framework, important confirmation and exceptions to our interpretation of scores would have escaped notice.

In our effort to follow Kane, we also considered methodological issues, including whether there was a need for IRT or other psychometric methods in test validation. As Marcoulides (2004) notes in his response to Kane, Kane takes a fragmentary approach that fails to appreciate the value added by coupling psychometric and non-psychometric approaches; we conclude, based on our own efforts and in keeping with Marcoulides’ critique, that the psychometric properties and validity of assessments must be considered in tandem for most test
validation efforts. We also observed, consonant with Haertel’s (2004) comments, that special attention must be paid to multiple-choice tests of professional skills, for the “unnatural formats” (p. 177) of these tests pose particular threat to the validity of interpretations of scores and require separate and very specific types of empirical exploration.

Finally, our own efforts suggest a slight modification to the structure of the validity argument itself. Specifically, Kane uses the phrase “assumptions and inferences” but makes no distinctions between the role of assumptions and inferences in either the formative or summative stages of test validation. In our view, assumptions and inferences are qualitatively different from the point of view of the test validation process; assumptions drive the substantive theory underlying the tests, whereas inferences are the potential consequences of the underlying theory and are what is empirically testable during test validation.

This paper consists of four major sections. The first section of this paper engages in a detailed critical examination of the validity argument approach, with an eye towards the generalizability of interpretative and validity arguments. The next two sections detail the motivation, conceptualization, and development of the MKT measures. The final section details the interpretive argument and outlines the remaining papers to follow in the series.

A CRITICAL EXAMINATION OF THE VALIDITY ARGUMENT APPROACH

While the validity argument approach is relatively new, as reflected in the 2004 Standards for Educational and Psychological Testing, it can be viewed as an extension of the construct validation approach advocated by Messick (1988, 1989). Consider the definition presented by Messick: “Validity is 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 or other modes of assessment.” (Messick, 1989) There are striking similarities between this statement and Kane’s argument-based approach.

However, there are a number of critical distinctions between Kane’s and earlier conceptions of construct validity. Most important is the explicit separation of validation into a formative stage, developing the interpretive argument, and a summative stage, assembling evidence to assess the interpretive argument. This separation focuses attention on the interpretive argument: it must be clearly stated, explicit, and detailed. Assumptions and inferences involved in the proposed interpretation must be spelled out, preferably in a step by step coherent chain; links should be either empirically verified, rejected, or verified with exception on the basis of evidence presented in the summative stage. Importantly, the process of accumulating and summarizing validity evidence should
be directed by the interpretive argument, with particular focus on the most questionable aspects of that argument. The separation of test validation into these two stages lends itself to viewing test validation as an iterative process.

While acknowledging the importance of the interpretive argument for organizing a wide variety of validity evidence, Schilling (2004) offered two main criticisms. First, despite Kane’s (2004a) desire to provide a methodology or technology for validation, this approach only provides a very general technology; the specific techniques employed are undetermined. Therefore, the argument-based approach is not prescriptive. Second, in attempting to present a single unified theory of validity, Kane ignores important distinctions among different types of tests, evidence, and stages of the validation process. Schilling argues that a better approach places these distinctions in the foreground and tailors methodologies to address these distinctions. In a rejoinder, Kane (2004b) rejected a specific form for the interpretive argument or specific validation techniques as too inflexible to be of any practical use.

Even though Kane rejects the notion of generalized interpretive arguments, a good place to begin developing such generalizations is by considering a specific example offered by Kane (2004a)—certification testing. Kane identified five major inferences constituting his interpretive argument for this specific example:

1. Evaluation of observed performance, yielding an observed score.
2. Generalization of the observed score to the expected score over the test domain.
3. Extrapolation from the test domain to the knowledge, skills, and judgment (KSJ) domain.
4. Extrapolation from the KSJ domain to the practice domain.
5. Decision about readiness for practice. (Kane, 2004a)

Below, we consider how general these arguments are, and by extension, whether they, in this or some other form, can be useful guides to practitioners engaged in test validation.

Clearly all tests must provide a justification from performance on the items to the generalization of an observed score. An important component is examining performance of examinees on individual items, deciding if the scoring rules for the individual items are correct and determining whether the skill being addressed or extraneous factors are responsible for correct performance on the individual items. Other issues that might be addressed here are whether multiple choice or constructed response item formats are appropriate for assessing the skill in question, addressing one of Haertel’s (2004) chief concerns in his critique of Kane’s (2004a) focus article. Test validation here proceeds at a molecular level, item by item. Therefore we term the assumptions and inferences elemental, because 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.

Kane’s second inference is specific to a generalizability approach (see Schilling, this issue), but his third inference is more universal; all tests must provide a justification that scales or subscales reflect the knowledge, traits, or skills the test is designed to measure. In addition, at this level one must provide justification for combining performance of the items into scale or subscales scores. Schilling (this issue) argues that the psychometric concept of essential unidimensionality—scale scores or subscale scores should be dominated by a single factor—should underlie both the specification of the structure of the knowledge, traits or skills domain being assessed and the empirical structure of the test itself. Therefore we term the set of assumptions and inferences addressed at this level structural assumptions and inferences; they incorporate Kane’s second and third inference.

Kane’s fourth inference is not universal; not all tests relate to practice. However, many of the issues involved in extrapolating to the practice domain in certification testing are generally involved in test validation. For example, if we consider driving tests, extrapolation to the practice domain involves demonstrating a relationship between successful performance on the test and external measures of successful performance in practice—accident rates, moving violations, etc. Similarly, a key component for evaluating the validity of the MKT measures or other measures of teacher knowledge would involve the relationship of test scores to measures of teaching and student performance. In both cases, relationships with external considerations and variables help establish the utility of the test. These considerations are often very specific to a particular test; in fact they probably constitute one of the more specific aspects of the interpretive argument. However, they are general in one respect: they place the test in context with respect to other important variables and specify the complex of relations between the test and these variables. We therefore use the term ecological assumptions and inferences to refer to this set of concerns, partly because they place the test in context and partly because, like ecological concerns in a biological context, the inter-relationships are often complex and specific to the particular test.

Finally there is Kane’s fifth inference. As was the case for the fourth inference, the decision about readiness for practice is specific to certification testing, but many of the concerns that would have to be addressed in this inference are general, such as the potential implications of the use of any test. These concerns have previously been encompassed under the term “consequential validity” (Moss, 1998). However, one can debate whether the potential intended and unintended consequences of the use of any test can always be assessed and whether such an evaluation is part of test validation generally. These issues are probably the most difficult to evaluate; to address them completely would be well beyond the scope of the current set of papers.
Specifying interpretive arguments as containing elemental, structural, and ecological assumptions suggests a toolkit of methodologies with which validation practitioners might do their work. Investigators can assess individual item performance—and thus the elemental assumption—via psychometric (i.e., item discrimination indices) and qualitative methods (e.g., think-alouds). Structural assumptions, likewise, can draw from psychometric traditions (e.g., factor analysis, assessing essential unidimensionality) and qualitative studies focused around how knowledge is organized and structured in individuals. Ecological assumptions can be assessed through correlational, quasi-experimental, or experimental studies; investigators can seek evidence both of positive correlations between theoretically related constructs and negative or null correlations between theoretically unrelated traits. Obviously, the methods we review above and actually use in this set of papers are not the only tools that could be used. However, one advantage of our organization of the validity argument is that it provides suggestions for the tools that could be used at each level of the analysis.

THE MKT MEASURES

We chose to explore the implementation of Kane’s validity argument approach in an applied setting, namely the Mathematical Knowledge for Teaching (MKT) scales developed for the Study of Instructional Improvement (SII) and Learning Mathematics for Teaching (LMT) projects at the University of Michigan. There are two reasons why this validation effort is important.

First, these measures are based on a well-developed but untested model of subject matter knowledge for teachers of mathematics, thus yielding a rich source of explicit assumptions and inferences for the interpretive argument. A number of our assumptions and inferences are different than typically found in achievement and certification tests, providing a good basis for determining the degree to which generalizations can be made with respect to the form of the interpretive argument, and the degree to which validity arguments are unique. The development and use of the MKT measures has yielded a wide variety of data, including responses of teachers, non-teachers, and mathematicians to the items comprising the scales, retrospective cognitive interviews of teachers, non-teachers, and mathematicians, classroom observations of teachers, and student gain scores on standardized achievement tests. These extensive data provide a means for assessing the varied assumptions and inferences arising from our model of mathematical knowledge teaching.

Second, measuring teachers’ mathematical knowledge is an important substantive issue in its own right. Recent years have seen an increased focus on describing the knowledge teachers use to teach mathematics (Ball, 1990; Ma, 1999) and determining how much knowledge individual teachers possess. The
scales developed at Michigan are based on work by Ball and others (Ball, 1990; Ball & Bass, 2003; Ma, 1999) that views mathematical knowledge for teaching as different from the types of mathematical knowledge common among mathematically competent adults—even mathematicians. For example, teachers not only need to perform basic computation for themselves, but also need to provide students with explanations for why particular procedures work, to diagnose student errors on those procedures, and to understand non-standard yet correct procedures. While this view has gained currency in the teacher knowledge literature, it has not been applied to construct reliable measures of teacher knowledge.

Motivation and Conceptualization

Research conducted over the past 30 years has clearly shown a connection between teachers’ knowledge and student achievement. Beginning with the Coleman report (1966), scholars have studied how what teachers know affects student gains (e.g., Hanushek, 1982; Mullens, Murnane, & Willett, 1996). In mathematics, however, the measures of teacher knowledge used in these “educational production function” studies were relatively simple and atheoretical: tests of mathematics basic skills, of the mathematics teachers are responsible for teaching students, and even of basic verbal ability. One often-drawn conclusion from this literature is that teachers’ general intelligence or verbal ability, as opposed to professionally-specific knowledge, leads to success in teaching (Walsh, 2001).

By contrast, a second stream of research has examined the knowledge that teachers use in working effectively with students. Beginning with Shulman (1986; see also Wilson, Shulman, & Richert, 1987), scholars in this area have sought to understand and document how discipline-specific professional knowledge can shape teachers’ classroom actions.

Consider, for instance, an elementary teacher providing her students instruction on the division of fractions. She clearly needs to be able to divide fractions for herself. However, she also needs mathematical knowledge specific to the work of teaching this topic—for instance, how to represent a problem such as $1\ 1/4 \div 1/2$ using a number line or area model, or how to explain the standard “invert and multiply” algorithm for division of fractions. She may also find knowledge of common student errors useful in instruction. For example, students often forget whether to invert the first or second term when solving a division of fractions problem; students will also improperly generalize from whole numbers to rationals, mistakenly assuming that division produces a smaller quotient than dividend. Finally, she may have special mathematical knowledge around the design of instructional tasks for students, either to introduce this topic or to remediate student errors. Many have come to see this knowledge as explicitly professional (Ball, Hill & Bass, 2005), different from the knowledge held by other mathematically competent adults.
The studies in which teachers’ mathematical knowledge is explored and defined (e.g., Ball, 1990; Borko et al., 1992; Ma, 1999) are mostly small-scale, and use only qualitative measures of teacher knowledge. Without measures that can reliably and validly measure teachers’ mathematical knowledge for teaching at scale, that is, for the potentially large number of teachers involved in an educational production function study, we cannot know whether and how this professional knowledge relates to students’ mathematics achievement. Without such studies, it is difficult to claim that the knowledge base this qualitative research describes is important and should be taught to beginning teachers. To understand what kind(s) of teacher knowledge—verbal facility, basic skills, or professionally specific knowledge—lead to student achievement, the field needed a measure that could be used at scale.

In 2000, an interdisciplinary group of researchers at the University of Michigan began to develop such measures as part of the Study of Instructional Improvement (SII). Later, item development and validation work was assumed by Learning Mathematics for Teaching (LMT), a National Science Foundation-supported project designed to support Math-Science Partnerships in the use of these measures to evaluate teacher learning.

Measure Development

Researchers began measures development by making a number of decisions regarding design. Most importantly, they wanted to develop measures of the mathematical knowledge used in teaching, not simply the content that teachers taught. This corresponds to the professional knowledge explored in the case study/qualitative literature described above. Second, given projections regarding the size of the teacher sample, researchers elected to use a multiple choice format. While existing measurement tools such as interviews, direct observations, and even discourse analyses had been fruitful in conceptualizing teacher knowledge, none were feasible for use in studies that would potentially include hundreds or even thousands of teachers. Third, because no map of “mathematical knowledge for teaching” existed in 2000, and anticipating such a map would take years to build, the authors declined to criterion reference the instruments. Finally, to match the focus of other SII instruments, the authors elected to write items in only three content areas: number, operations, and patterns/functions/algebra. The first two comprise the bulk of mathematics lessons taught in elementary schools (Rowan, Hayes, & Harrison, 2004), while the last has been a focus of reform efforts in K-6 mathematics education.

Once the content domains had been chosen, development of the measures focused on specifying the types of professional knowledge researchers considered important to capture. Initial work centered around two constructs: content knowledge (CK), and knowledge of content and students (KCS). Below, we
define these two domains and provide examples of items written to represent them. Knowledge of content and teaching (KCT) has been conceptualized but not yet measured.

We defined the CK domain as consisting of subject matter knowledge—in this case, mathematics itself. Within CK, however, we hypothesized two possible subcategories. “Common content knowledge” (CCK) refers to commonly held mathematics knowledge—that which crosses disciplines and exists in the public domain. By contrast, “specialized content knowledge” (SCK) is mathematical knowledge specific to the profession of teaching. The latter category consists of mathematical tasks such as representing numbers and operations with pictures or manipulatives, examining and generalizing from non-standard solution methods, and providing explanations for mathematical ideas or procedures. These are all mathematical tasks, but often not encountered by those who use mathematics in everyday tasks or other professions. Two items in the Appendix provide examples of common and specialized content knowledge. Item 1 is an example of a “common” mathematics knowledge task; many adults and all mathematicians should be able to answer this item correctly. Item 2 is a specialized knowledge task. Teachers must determine which picture does not map onto the number sentence $1 \frac{1}{2} \div 2 \frac{2}{3} = 1$; a key sensitivity is to the representation of $1 \frac{1}{2}$, which must be on the same-sized objects. Using both a rectangle and circle, as in example C, is like representing this number as one apple and $1/2$ an orange. This knowledge may not be apparent to those who do not teach mathematics to students, yet it is purely mathematical—it is specialized to the work of teaching.

The KCS domain was defined by stipulating that respondents would ideally use knowledge of students’ thinking around particular mathematics topics, rather than purely their own mathematical knowledge, test-taking skills, or other processes, to arrive at their answer. Because this domain is an amalgam of knowledge of content and knowledge of students, teachers might also invoke mathematical knowledge or engage in mathematical reasoning in order to interpret students’ thinking around these topics. However, respondents must use knowledge of students above and beyond purely mathematical knowledge. The definition of knowledge of students’ learning of mathematics in actual item-writing included not only findings from the research literature (e.g., on students’ understanding of the equals sign; Carpenter, Franke & Levi, 2003), but also the knowledge observant teachers might glean from working with students, but that has not been codified in the literature (e.g., what third graders are able to do, mathematically). Two items in Appendix 1 provide examples from the KCS domain. Item 3 asks teachers to anticipate what mistakes students might make while representing the number 16 with base ten blocks. Research in mathematics education has demonstrated that students are likely to mistakenly use one cube and six cubes, or 16 cubes in a pile, and a rod and six cubes is the correct answer. Thus seven cubes in a pile is the least likely. Item 4 requires that teachers know a common student error in ordering decimals—treating decimals as if
they were whole numbers. Students making this error would nevertheless correctly place the numbers in (a) and (b). Only choice (c) would allow teachers to detect this problem.

After specifying the measurement domains, project members wrote and piloted items; a total of six elementary forms and three middle school forms have been piloted to date. Data collection to enable validity analyses were begun soon after, and include interviews and videotaped classroom observations. Together, these data sources form the basis for the studies described here.

**INTERPRETIVE ARGUMENT FOR THE MKT MEASURES**

Based on our project’s conceptual orientation toward mathematical knowledge for teaching, and on the measures’ potential uses in the field of research, we developed an interpretive argument during the formative stage of our validation process. The interpretive argument followed the distinctions we made above between key types of validity, and consists of the following assumptions and inferences:

1. Elemental assumption: The items reflect teachers’ mathematical knowledge for teaching and not extraneous factors such as test taking strategies or idiosyncratic aspects of the items (e.g., flaws in items).
   A. Inference: Teachers’ reasoning for a particular item will be consistent with the multiple choice answer they selected.

2. Structural assumption: The domain of mathematical knowledge for teaching can be distinguished by both subject matter area (e.g., number and operations, algebra) and the types of knowledge deployed by teachers. The latter types include the following: content knowledge (CK), which contains both common content knowledge (CCK), or knowledge that is common to many disciplines and the public at large and specialized content knowledge (SCK) or knowledge specific to the work of teaching; and knowledge of content and students (KCS), or knowledge concerning students’ thinking around particular mathematical topics. Implications of this include:
   A. Inference: Items will reflect this organization with respect to both subject matter and types of knowledge in the sense that items reflecting the same subject matters and types of knowledge will have stronger inter-item correlations than items that differ in one or both of these categories. This will result in the appearance of multiple factors in an item factor analysis.
B. Inference: Teachers can be reliably distinguished by unidimensional scores reflecting this organization by subject matter and types of knowledge. These scores are invariant with respect to different samples of items used to construct the scores.

C. Inference: Teachers will tend to answer most problems (except those representing CCK) with knowledge specific to the work of teaching. Non-teachers will rely on test-taking skills, mathematical reasoning, or other means to answer these items.

D. Inference: Teachers’ reasoning for a particular item will reflect the type of reasoning (either CK or KCS) that the item was designed to reference.

3. Ecological assumption: The measures capture the content knowledge that teachers need to teach mathematics effectively to students.

A. Inference: Higher scores on the scales derived from these measures are positively related to higher-quality mathematics instruction.

B. Inference: Higher scores are positively related to improved student learning.

In the papers that follow we use multiple sources of evidence to evaluate these assumptions and inferences. Hill, Dean, and Goffney (this issue) used cognitive interviews to address the elemental assumption, determining whether respondents’ reasoning corresponds with their answers; it also addresses the structural assumption by determining whether the reasoning used reflects the types of knowledge hypothesized to exist as part of MKT. Schilling (this issue) also addresses the structural assumption, by using unidimensional and multidimensional item response theory to map the item domain to the hypothetical structure of MKT. Hill et al. (this issue) assesses the ecological assumption and interpretations, including the relationship of content knowledge measures to student outcomes and mathematics teaching practice. Finally, Schilling, Blunk, and Hill (this issue) attempt to draw some final conclusions regarding the MKT measures, detailing what we have learned, its implications for these measures and its implications for practical implementations of the validity argument approach.

ACKNOWLEDGMENTS

Research for this paper was supported by NSF grants REC-0207649, EHR-0233456, and EHR-0335411. The authors would like to thank Deborah Ball, Hyman Bass, Merrie Blunk, Catherine Brach, Seán Delaney, Jennifer Lewis, Laurie Sleep, Mark Thames, and Deborah Zopf for their help in developing aspects of this paper. Errors remain the property of the authors.