Test Validation and the MKT Measures: Generalizations and Conclusions

Stephen G. Schilling a, Merrie Blunk a, Heather C. Hill b

a University of Michigan, Ann Arbor, Michigan, USA
b Harvard Graduate School of Education, Cambridge, Massachusetts, USA

Online Publication Date: 01 August 2007


To link to this article: DOI: 10.1080/15366360701487146
URL: http://dx.doi.org/10.1080/15366360701487146

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Department of Education to the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania (Grant # OERI-R308A60003) and the Center for the Study of Teaching and Policy at the University of Washington (Grant # OERI-R308B70003); and to grants to the University of Michigan by the National Science Foundation (IERI/ REC-9979863 & REC-0129421, REC-0207649, EHR-0233456, and EHR-0335411), the William and Flora Hewlett Foundation and the Atlantic Philanthropies. Opinions expressed in this paper are those of the authors, and do not reflect the views of the U.S. Department of Education, the National Science Foundation, the William and Flora Hewlett Foundation or the Atlantic Philanthropies.

Test Validation and the MKT Measures: Generalizations and Conclusions

Stephen G. Schilling and Merrie Blunk
University of Michigan, Ann Arbor, Michigan

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

This series of papers had two main objectives: (1) Use the validity argument approach to critically assess the assumptions underlying the MKT scales; (2) Use this experience to critically assess the validity argument approach. We argue that understanding the validity argument approach arises from real world applications to specific examples. Engaging in these applications suggests possible methodologies for the validity argument approach.

Schilling and Hill presented the validity argument approach, introduced the MKT measures, and developed an interpretive argument. Drawing on the work of Kane, we specified an interpretive argument that contained three types of assumptions and inferences: an elemental assumption—that answers to individual items correspond to their reasoning for those items; a structural assumption—that the organization of items reflects the types of MKT employed

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
by teachers, specifically common content knowledge (CCK), specialized content knowledge (SCK), and knowledge of content and students (KCS); an ecological assumption—that the items reflect the knowledge that teachers need to teach mathematics, and scores are therefore related to the mathematical quality of instruction and improved student learning. Hill, Dean, and Goffney investigated the elemental structural assumptions, examining interview transcripts to determine respondents’ reasons underlying answers to the items. Schilling examined the structural assumption, illustrating the importance of specifying a test domain structure consisting of unidimensional components and showing how multidimensional and unidimensional IRT models can be used to test this structure. Hill et al. investigated the ecological assumption, examining the correlation between teachers’ performance on our instruments and classroom mathematics instruction and student achievement.

In this paper, we complete the summative stage of the validity argument approach, then use our experiences to reflect on the validity argument as a method. We begin by evaluating the inferences and assumptions of the interpretive argument for the MKT measures. Then we examine both the form and the structure of the interpretive argument for the MKT measures with an eye to generalizations that can be made to other efforts to construct interpretive arguments. Finally we will attempt to draw some reasonable generalizations and conclusions concerning the summative stage.

EVALUATION OF THE MKT INTERPRETIVE ARGUMENT

We examined the MKT interpretive argument in a number of studies, each employing different methods. The challenge now is to integrate this varied data, drawing substantive conclusions concerning the inferences contained in the interpretive argument and, if necessary, reformulating that argument in light of this evidence. What have we learned in this process about the MKT measures? Our findings vary by construct, and this is how we organize the discussion below.

The content knowledge items and measures, as a whole, behave in approximately the manner we expected they would. Cognitive interviews support the elemental assumption, that a teacher’s reasoning for a particular item is consistent with the multiple choice answer she selects (Hill, Dean, & Goffney, this issue). Evidence supports both inferences in the ecological assumption: higher scores on the measures are related to higher-quality mathematics instruction, where higher-quality is defined by examining the mathematics that occurs in the classroom (Hill, Ball, Blunk, Goffney, & Rowan, this issue); higher teacher scores on the measures are also positively related to gains in student learning (Hill, Rowan, & Ball, 2005).
One revision to the interpretation of content knowledge scores concerns our assertion that content knowledge contains two sub-constructs, common (CCK) and specialized content knowledge (SCK), that are separable from one another as constructs. Results in Schilling (this issue) both showed the CCK and SCK items did not differentiate themselves in exploratory factor analysis and that “content knowledge” scales composed of both CCK and SCK items do not meet criteria for essential unidimensionality. Item vector plots suggested that SCK items based on nonstandard methods and some representations behaved differently than common content knowledge (CCK) items. Interviews with mathematicians also suggested that while most had little trouble answering our problems correctly, the areas that proved problematic included tasks of teaching we would consider specialized—e.g., flexibility with nonstandard approaches and explaining mathematical ideas to learners. It is noteworthy that these SCK items were also identified by multidimensional IRT methods as being oriented in a different direction as the CCK items. However, not all representation and explanation items consistently behaved as SCK.

In the face of this evidence, our strategy, as we discuss below, is not to abandon this element of the structural assumption, but to continue to build our notion of what constitutes specialized knowledge for teaching, refining the specification of this construct and the items written to represent it. Our new definition, for instance, will start with the elements of SCK indicated here—nonstandard methods and providing “decompressed” explanations and representations for student use; we will then build additional categories within SCK based on our validation work observing classroom mathematics. We will also need to reformulate our observable implications of theory, perhaps arguing for a more explicit testing of “expert” and “novice” teachers to see what elements of mathematical knowledge for teaching are purely mathematical, and which come from either professional education or on-the-job experience.

More problems, and more revisions to our assumptions, were revealed by evidence regarding the knowledge of content and students (KCS) items. While our structural analysis showed a KCS factor was present in our measures, some items we designed to measure KCS appeared on the CK factor; even more problematic, the KCS items did not define an essentially undimensional scale without deleting items. Interviews confirmed that many respondents used only mathematical reasoning to answer KCS items. Interviews also suggested that some respondents used test-taking strategies and guesswork to answer these items, violating the elemental inference. This evidence, in turn, helps explain why we have found that these items are typically less reliable than the content knowledge items (Hill, Schilling, & Ball, 2004). This lower reliability prevented the use of KCS items in our study relating teacher knowledge to student achievement.

In light of this evidence, we have substantially revised our theory regarding knowledge of content and students, and may also need to revise our measurement
strategy, as well. When we began developing items in this domain, we hypothesized that teachers’ knowledge of students existed separately from their mathematical knowledge and reasoning ability. We thought of such knowledge as “declarative,” or factual knowledge teachers have of student learning. Results from these validation studies, however, suggest that this “knowledge” may actually contain both elements of mathematical reasoning and knowledge of students and their mathematical trajectories. From this standpoint, KCS is less declarative knowledge than a kind of close reasoning in which teachers engage, flexibly, about students and their work. This reasoning is aided by teachers’ experiences with previous student learning and mistakes. From a measurement standpoint, this complicates matters greatly, for this domain is theoretically multidimensional, containing both facts and reasoning.

Along with this theoretical redefinition, we are also seeking to improve the actual items used in this category, moving toward ones that better capture the knowledge teachers have of students’ mathematical learning. This may take a reorientation of our measurement strategy away from multiple-choice (which tend to be quite easy for teachers and nonteachers alike) toward open-ended responses.

GENERALIZING THE FORMATIVE STAGE: THE STRUCTURE OF THE INTERPRETIVE ARGUMENT

In his 2004 article, Kane formulated an interpretative argument that contained five major types of inferences about certification tests: (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; and (5) Decision about readiness for practice. Based on our work, we argue for two critical amendments to Kane’s formulation. First, we advocate separating the interpretive argument into assumptions and inferences. Second, while Kane’s inferences are essentially hierarchical, that is, stated in terms of generalizations from lower to higher levels, our general structure separates the assumptions and inferences into elemental, structural, and ecological components and focuses on establishing connections for each component. We believe interpretative arguments should contain at least these three components.

Consider the separation of each component of the interpretive argument into assumptions and inferences. This feature of the argument is not well developed in Kane’s writing: his discussion of informal reasoning and the certification example (Kane, 2004a) focuses on inferences while his discussion of the interpretive argument (2001, 2004a) mentions both assumptions and inferences.
However, their relationships both to each other and to the validation argument as a whole were underspecified. In our validation argument, it became evident that both inferences and assumptions need to be explicitly stated, and that they differ with respect to the ease with which they can and should be refuted based on the available evidence. Inferences are eminently falsifiable; properly conducted studies which fail to yield support will cause the inference to be discarded, severely weakened or modified in light of the empirical data. However, discarding unsupported inferences does not necessarily mean discarding the assumptions to which they refer. Rather, inferences arising from assumptions can be replaced or modified so that they become consistent with both the assumptions and empirical data.

A case in point is our structural assumption and its inferences concerning SCK. The notion that teachers’ mathematical content knowledge is specific to the work of teaching is central to the design and development of the MKT scales. If true, this suggests that SCK should be distinguishable from CCK. One way this distinction should manifest itself empirically is in stronger inter-item correlations among SCK items than between SCK and CCK items, resulting in separate factors in the item factor analyses. But the actual analyses revealed that this was not the case for most SCK items. However, we need not abandon our structural assumption if this inference can be modified or replaced and remain consistent with both the assumption and related empirical evidence.

For example, one possible explanation for the failure of the SCK items to load on a separate factor in the item factor analysis is the relationship of the SCK construct to the CCK construct in the samples used for the analysis. It might be that for a homogeneous population of teachers, some aspects of knowledge tapped by the SCK items is strongly associated with CCK, but that in a more heterogeneous population this relationship might not hold. The key to understanding this is to understand that lack of unidimensionality is the result of stronger association between certain collections of items and that association can be the result of population heterogeneity.

For example, all the SCK items concerning nonstandard methods appeared on the second factor in the factor analyses. One way in which stronger association between these types of items could occur is if there is a subpopulation of teachers who have been taught to recognize and work with nonstandard algorithms and if there is another subpopulation that has not been given this opportunity. These subpopulations will be reflected in the diagonals of two-way contingency tables of association between these items, thereby increasing the association between these items. Conversely, if the subpopulation is homogeneous with respect to this opportunity to work with nonstandard algorithms, this extra source of bivariate association between the items is not present and any covariation between these items is due to the common content knowledge in the items.
If this is the case, whether SCK shows up at all as a separate factor in factor analysis, or whether it shows up for only certain subsets of items will be very population specific. In fact, we have subsequently found that SCK items are very inconsistent across different samples with respect to loading on a separate factor. But this does not necessarily refute Ball and Bass’ view that SCK is an important additional component in mathematical knowledge for teaching. Rather we need to revise the inference that it will show up as a separate essentially unidimensional component in item factor analyses. It may be that we need to use other methods to investigate SCK. For example, investigations of subpopulations that have less opportunity to learn aspects of SCK (e.g., new teachers, mathematicians, other mathematically competent adults) might reveal differential item functioning on SCK items compared to the general population of teachers. For a full discussion of the relationship of multidimensionality and differential item functioning see Shealy & Stout, (1993).

A second element of the structure of this interpretive argument worth noting is the nature of the assumptions and inferences. They can be generally characterized as follows:

Elemental: Assumptions and inferences concerning the internal structure of items—consistency of items with subjects’ knowledge.
Structural: Assumptions and inferences concerning the internal structure of test—the consistency of the structure of the test with the structure of the test domain.
Ecological: Assumptions and inferences concerning the external structure of the test—the configuration of relationships of the test scales with external variables.

The assumptions and inferences, while specific to the MKT measures, are general in terms of their characterization. We contend that interpretive arguments should always consist of these three elements, for the reasons outlined below.

Elemental assumptions and inferences are often overlooked but are a prerequisite for the other two assumptions. Here we are operating at the item level, determining subjects’ reasoning on the task, their reasons for generating a particular answer, and whether task reasoning and answer generation are consistent. If consistency cannot be established, one cannot determine what the items and therefore the test are assessing. Lack of consistency implies construct-irrelevant variance.

Structural assumptions and inferences are universal to all test construction—one must delineate the domain of knowledge, skills, and judgment to be assessed, establish its structure, and detail the correspondence between items, scales and the domain. While ecological assumptions and inferences establish why the test is important, structural assumptions and inferences describe how the
test and its scales represent the core elements in the intended domains of measurement. Verifying this construction is a prerequisite to the ecological assumption.

Our specification of the structural assumptions differs from how Kane attempts to relate observed score to the KSJ domain in that we consider the test domain and the KSJ domain jointly—there is no extrapolation from the test domain to the KSJ domain. Also, following the lead of the Schilling (this issue) article, we emphasize the utility of a specification of the test domain in terms of essentially unidimensional components and establishing the correspondence of these components with the theoretical specification of the domain presented in the assumptions and inferences. If the domain cannot be resolved into essentially unidimensional components that correspond to the theoretical constructs, a re-specification of the test domain is required. It is likely that re-specified constructs are under-represented in terms of number of items for the test. For example, re-specifying the SCK construct as consisting of those items that did not fall along the validity sector defined by the CCK items resulted in too few items to reliably measure that construct. Construct under-representation was identified by Kane (2004a) as one of the two primary threats to the extrapolation from the test domain to the KSJ domain. The other, construct irrelevant variance, is assessed by the elemental assumption.

Ecological assumptions and inferences have considerable overlap with Kane’s extrapolation from the KSJ domain to the practice domain; in our case it is a direct statement about the relationship of the test to the teachers’ effectiveness in teaching mathematics. However, Kane’s formulation assumes that the objective of the test is to represent the entirety of the practice domain with the test, or at least those elements of the practice domain that are critical for success. Restrictions on the knowledge or skills assessed by the test are the result of limitations of the test format itself. For example, teaching mathematics effectively involves maintaining class order and discipline, interacting with parents and administrators, and other such nonmathematical skills, which might be difficult to assess in an objective test, and which are likely to be separate from the construct described by MKT. The difficulty with a specification that attempts to generalize from the test to practice is that, even in Kane’s certification example, a test will always be viewed as inadequate or limited.

Some may argue that the three-part structure we have imposed on the interpretive argument is not necessary, and need not be applied in the general case. To respond, we examine the potential consequences for test validity of failure to assess one or more assumption(s) and inference(s). Failure to assess ecological assumptions and inferences implies a lack of justification for the use of measures. If test scores have no relationship to external variables, then the test likely has little import. For example, in Kane’s certification example, failure to establish the relationship of the test to practice would necessarily call into question the
usefulness of the test. Similarly, if the MKT measures have no relationship to teaching mathematics or student achievement, why use the measures?

Failure to critically assess elemental assumptions and inferences leads to uncertainty concerning whether performance on an item reflects performance on the presented task or some other aspect of the item, such as reading comprehension or test taking strategies. For example, if options for a multiple choice item can be eliminated absent subject matter knowledge, correct performance on that item is the result of test-taking skill. If this occurs for a large number of items, test performance is a function of test-taking skills, rendering the scales ineffective for evaluating teacher learning.

Failure to critically assess structural assumptions and inferences leads to a lack of knowledge concerning the structure of the test, the structure of the test domain, and the correspondence of these structures to the implicit structure of the KSJ domain. As discussed in Schilling (this issue), if the test and test domain are highly structured and unidimensional and we can establish consistency of the test to the test domain, we have a basis for interpreting reliability and generalizability coefficients. Otherwise interpretations of reliability and generalizability occur in a vacuum.

Test developers and consumers typically assume unidimensional domains for test scales or subscales. Developers apply psychometric models to the scales or subscales corresponding to different aspects of their test domains and compute reliabilities and other test indices. However, if the test scales or subscales are not unidimensional, it becomes difficult to establish correspondence between scores on different forms of a particular test, much less between different tests purporting to measure the same aspect of the test domain. For example, we were able to apply IRT models, compute reliabilities, and estimate scales scores for CK, CCK, SCK, and KCS scales. But effectively equating forms was only possible for the CCK and CK scales because only they were essentially unidimensional. Conversely, a broad unstructured test domain leads to the oft-noted dilemma concerning the tradeoff between reliability and generalizability: narrowly defined test domains lend themselves to high reliability at the expense of generalizability; broadly defined test domains lend themselves to generalizability but high reliability is more difficult to achieve.

GENERALIZING THE SUMMATIVE STAGE

Based on the experiences described above, we contend that the summative stage of validity arguments should have, as a rule, some common features. First, evidence for the validity argument should employ multiple methodologies, both quantitative and qualitative. This is particularly true for ecological and structural assumptions. For example, in our evaluation of the structural inferences, the interview data was invaluable. Absent this data, we would have been unsure how
to interpret the multidimensionality of the KCS items; using this data, we saw that some respondents answered items correctly by using mathematical reasoning alone. Similarly, our interviews with mathematicians indicated that individuals not engaged in the work of teaching often had difficulty with SCK items. In both cases, qualitative data provided a level of detail beyond that supplied by the statistical analyses alone.

Second, our work reinforces Kane’s assertion about the importance of a critical approach when collecting and examining validity evidence. To use Kane’s words, “A critic who questions the interpretive argument is likely to be supplied with the appropriate warrants. The role of the validator is to be an especially persistent critic.” What are the implications of this approach in the context of our analyses? In order to meet Kane’s criteria, we cannot opportunistically use results from the analysis to support our interpretive argument.

For example, consider alternative conclusions that could have been drawn from the results of our analyses for the KCS items. We could have effectively argued that the plots of the factor loadings led to two groupings of the items—CK and KCS—and omitted the detailed exploration of the multidimensionality of the data. A Rasch model fit the KCS items adequately and provided acceptable levels of reliability (Hill, Ball & Schilling, in press). This would provide support for inference 2B for KCS. However, only the detailed multivariate exploration of the data and the equating results indicated flaws for this inference.

We are less inclined to be prescriptive with respect to specific methodologies and techniques, because the particular methods and techniques chosen are necessarily tied to the nature of the inferences being investigated. Rather, our application provides a sampling of the variety of methods and techniques that might be used. However, we do advocate the use of an explicit psychometric theory and psychometric models in assessing structural inferences. We do not view psychometric modeling, reliability, and validity to be separate enterprises, but rather part of a continuum of interconnected objectives oriented to determining the meaning of test scores. In our case, we found that full information item factor analysis (Bock & Lieberman, 1970; Bock, et al, 1988; Bock & Schilling, 1997; Schilling & Bock, 2005) provided us with an effective means for investigating the structure of the test domain. However, this is not the only method that can be generally employed, even within the context of IRT. For example, Embretson and Gorin (2001) demonstrated how IRT modeling employing cognitive principles can assess construct validity.

Prescriptions concerning techniques to investigate ecological and elemental assumptions are even less specific. One reason is, as Kane (2004a) noted, these assumptions and inferences are less amenable to precise analysis through mathematical modeling than structural assumptions. Therefore, the objective is to develop a toolkit of possible methods that could be adapted to project-specific assumptions.
Possible components of the ecological toolkits might include some subset of experimental, quasi-experimental, correlational, and observation studies and the host of quantitative, statistical, and qualitative methods that have been developed for each of these types of studies. At this stage in the MKT investigation, ecological assumptions and inferences were assessed by videotape studies of classrooms using subjective ratings of teachers’ performance and a correlational study using hierarchical linear models. However, future work might involve either experimental or quasi-experimental studies.

One component of the elemental toolkit employed in our investigation was retrospective cognitive interviews, but as noted in Hill, Dean, and Goffney (this issue) we could have also employed a think-aloud protocol. From the perspective of psychometric methods, it might be fruitful to apply Bock’s (1972) nominal response model for multiple-choice items to determine if incorrect responses are associated with the knowledge or skill being measured by other items.

CONCLUSION

Our application of the validity argument approach has been motivated by a dual desire to better understand the MKT measures and provide guidance to others concerning general principles about argument-based test validation. Like Kane’s argument-based approach, we view our work as part of an iterative process, with much still be left done on both fronts.

Our understanding of the MKT measures and the content knowledge necessary to teach mathematics has clearly been enhanced by this process. As we noted at the outset of the paper, measure development served as a test of the theory that teachers have specific and special knowledge of mathematics. We can say that the items, particularly the common and specialized content knowledge items, work at some broad level to discriminate between teachers who have and do not have mathematical knowledge, and to predict instruction and student learning. This, in itself, is one form of validation; representing an idea in actual items, then seeing those items work on at least a basic level to measure teachers is a success. Whether these items measure knowledge that is unique to teaching is still a question for our own development and for the field of mathematics education more broadly.

Also, because we are still developing a theory of teacher knowledge, the effort invested in validation has proven useful in more ways than reporting that these measures work. Viewing videotapes of actual teaching, examining results from factor analysis, and discussing our items with mathematicians have all contributed toward our own evolving notion of MKT and its component parts. These sources of evidence have also helped develop new items, which will be used in subsequent rounds of measure construction and validation.
Our understanding of the validity argument approach has also been enhanced by our experience. We are prescriptive in the sense that we specify some necessary first steps when employing this approach. However, we recognize that in any particular application there will be specific considerations that go beyond our prescriptions. More importantly, we have provided an application of the validity argument approach with a degree of detail that had been absent in the validation literature. The principles and techniques we have employed are useful, and provide a degree of guidance missing in Kane’s original formulation. However, much future work needs to be done to explicate the full range of principles and techniques needed in test validation using the validity argument approach.

REFERENCES


