ISSN 1536-6367
MEASUREMENT
Interdisciplinary Research and Perspectives
Editors
Mark Wilson
Paul De Boeck
Pamela Moss
LEA Lawrence Erlbaum Associates, Publishers

Measurement: Interdisciplinary Research & Perspective
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653679

The Role of Psychometric Modeling in Test Validation: An Application of Multidimensional Item Response Theory

Stephen G. Schilling [a]
[a] University of Michigan, Ann Arbor, Michigan, USA

Online Publication Date: 01 August 2007

To cite this Article Schilling, Stephen G.(2007)'The Role of Psychometric Modeling in Test Validation: An Application of Multidimensional Item Response Theory',Measurement: Interdisciplinary Research & Perspective,5:2,93 — 106

To link to this Article: DOI: 10.1080/15366360701487021

URL: http://dx.doi.org/10.1080/15366360701487021

The Role of Psychometric Modeling in Test Validation: An Application of Multidimensional Item Response Theory

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

One of the key challenges facing psychometrics as a discipline is demonstrating its relevance with regards to substantive issues in educational and psychological research. One problem is that psychometric modeling has long been considered part of reliability analysis, which has traditionally been considered separately from test validation. Kane (2004a) explicitly made this distinction:

> "... validity has proven difficult to pin down ... in part because interpretations and uses are less amenable to precise analysis ... than other aspects of measurement, particularly scaling and reliability."

However, this separation can be viewed as artificial (Marcoulides, 2004) and we believe that test validation must *necessarily* employ psychometric modeling to investigate key assumptions and inferences. Such an investigation intimately connects psychometric modeling to substantive concerns and provides a gateway for the relevance of psychometrics in educational and psychological research.

In this paper we examine the role of item response theory (IRT), particularly multidimensional item response theory (MIRT) in test validation from a validity argument perspective. Our conceptualization of the interpretive argument differs from Kane in that it combines his second and third inference—generalization from the test score to the expected score over the test domain and extrapolating from test domain to the knowledge/skill/judgment domain—into a general

Research in this paper was supported by NSF grants REC-0207649, EHR-0233456, and EHR-0335411. The author would like to thank the LMT research project for helpful feedback and advice.

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

category called "structural assumptions and inferences." We base this move on our view that the test and knowledge, skills, and judgment (KSJ) domains do not usually exist extant; rather, they are typically constructed according to theory that specifies a structure to the test and KSJ domains. Given this, MIRT can help validate the proposed theoretical structure of the test, and thus provide evidence regarding the validity of test scores.

Specifically, we examine the first two inferences from our structural assumption:

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:

1. 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.
2. 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.

The outline of this paper is as follows. In the next section we provide justification for these structural assumptions and interpretations, taking care to describe the role we believe they should play in any interpretive argument. Then we show how MIRT methods using full-information item factor analysis (Bock & Aitken, 1981; Bock, Gibbons, & Muraki, 1988; Schilling and Bock, 2005), multidimensional item difficulties (Reckase, 1985), and validity sectors (Ackerman, 1994) can be used to establish essentially unidimensional scales corresponding to the theoretical structure of the test domain. We then apply these methods to assess the three components of MKT. Finally, we discuss the implications of our findings for our theory of MKT and for psychometric theory generally.

DELINEATING THE STRUCTURE OF THE TEST DOMAIN: KEY CONCEPTS

Our structural assumption originates from a view that validation practitioners should specify a proposed structure for the test and empirically assess whether the test matches that structure. This differs from the approach used by Kane (2004a, 2004b), which arises from a generalizability theory perspective—first generalizing from the test score to the test domain, then extrapolation from the test domain to what Kane terms the knowledge/skill/judgment (KSJ) domain. In Kane's second inference, generalizing from the test score to the test domain, test items must be either a random or representative sample from a larger domain and the number of items in the test must be large enough to produce a dependable estimate of the expected score on the test domain.

The difficulty with the generalizability theory perspective, particularly in the context of test validation, is that it assumes a defined test domain from which random or representative sampling is possible. However, this is only true in certain very specific cases, such as spelling tests, where the domain to sample from (a dictionary) is easily defined. Generally the test domain is not so easily defined. Rather it must be constructed by first specifying a structure of the domain, constructing items to represent aspects of that structure, and then defining scales or subscales on the basis of establishing psychometric rules for scoring based upon that structure. For example, in reading tests for first grade elementary school children, the domain can be defined as consisting of items concerning letter sound correspondence, morphology, vocabulary, decoding skills, word reading, and reading comprehension. One possible structure for such a test could have the letter sound correspondence, morphology, and decoding skills items constituting one scale—word analysis, and the word reading, vocabulary, and reading comprehension items constituting another scale—reading comprehension.

Both the reality that most test construction proceeds via the construction of a content or domain map as well as test consumers' interest in performance on related but distinct areas provide an argument for the role of psychometrics in test validation. At a minimum, validation efforts should empirically analyze the assumptions about test structure made during test construction.

We also argue that for an assessment to be a good summary of the performance of individuals, its scales need to be essentially unidimensional (Stout, 1987, 1990). If the scale and the items that comprise it are essentially unidimensional, scores predict performance on the items. It is then reasonable to summarize performance on the collection of items by a single score. If the scale is not essentially unidimensional, it means that the scale score does not predict performance for some subset of items—there is additional information for this subset of items that is not being reflected in the total score. If the subset is large enough, it can alter scores to the degree that it is impossible to get consistency across different forms. To use Kane's (2004a) terminology, essential unidimensionality provides the "warrant" or justification for assigning a single score to examinees' responses on a collection of items.

We can contrast essential unidimensionality with strict unidimensionality. Strict unidimensionality says that any association between items is due to a single factor; after fitting a single factor model to test data the residual correlations between any two items should be zero. Essential unidimensionality relaxes this assumption; it only requires that the average residual inter-item correlations after fitting a one-dimensional IRT model approach zero as the number of items increases. It allows for minor deviations from unidimensionality that constitute spurious factors, but still stresses the fact that any test or scale should be dominated by a single factor or dimension.

This enables us to have a more practical and realistic definition of unidimensionality, one that is related to the concept we are trying to measure. For example, in reading comprehension tests, essential unidimensionality is not concerned with individual passage effects because the residual correlations due to passage effects go to zero as the number of test items increase. Essential unidimensionality only addresses whether the questions across passages are measuring the same unidimensional trait—the ability to read and comprehend text.

Given the importance of specifying the test domain in terms of essentially unidimensional components, we next turn to how to explore the dimensionality of tests and verify a theoretically-specified structure in terms of essentially unidimensional components.

METHODS FOR ASSESSING MULTIDIMENSIONALITY AND DETERMINING ESSENTIALLY UNIDIMENSIONAL TEST COMPONENTS

One general method for exploring test dimensionality, establishing essentially unidimensional scales, and connecting these scales to a specified theoretical structure is exploratory full-information item factor analysis and MIRT. Full-information item factor analysis has a number of advantages when compared to other means of factor analysis, such as general least squares (GLS) as implemented in MPLUS, including better recovery of factor loadings in the presence of varying difficulty levels and more accurate and precise likelihood ratio tests of the number of factors (Schilling & Bock, 2005). Sometimes the statistical tests and rotated factor loadings are simple to interpret. However, most often we need to go beyond the simple statistical tests and tables of rotated factor loadings by employing a variety of tools, including graphical analysis measures of fit and residual analysis to determine essentially unidimensional components.

The situation is akin to regression analysis or other statistical procedures where a variety of tools, including tests comparing models of different orders, measures of fit such as adjusted R-squared, graphical analysis, and residual analysis are used to help researchers make important substantive decisions. As is the case with regression analysis, these sources of information must be combined using judgment. As an aid to judgment, we often employ rules of thumb or heuristics at key decision points.

The graphical tools we use are what Reckase and Ackerman (Ackerman, 1994; Reckase, 1985; Reckase & McKinley, 1991) call item vector plots. Item vector plots are scatterplots of two-dimensional item difficulties (Reckase, 1985) with respect to an orthogonal factor analysis solution with two factors. However, vectors with length proportional to the multidimensional discrimination (Reckase, 1985) are placed in the plot originating from the two-dimensional item difficulties and oriented in the direction indicated by the two-factor orthogonal solution. The advantage of this type of plot is that it allows us to determine if a collection of items approximately determines a line in two-dimensional space. If so, it is a good indication that the set is essentially unidimensional. One way of measuring this is the item sector width (Ackerman, 1994)—the variation in the item vector angles. A useful heuristic for determining essential unidimensionality is that most of the item vectors should lie in a sector of approximately 30 degrees or less.

In addition we use two other summary statistics to help us determine essential unidimensionality. The first of these, based on the Akaike Information Criterion, is just the chi-square statistics for comparing a 1-factor to a 2-factor model divided by two times the degrees of freedom—we call this the AIC goodness of fit index (AIC GFI) for unidimensionality. Values less than one indicate essential unidimensionality, while values greater than one indicate at least two dimensions are needed. A final measure of unidimensionality is simply the root mean-squared residual correlations of the fit for a one-dimensional model. A good rule of thumb for determining essential unidimensionality is that the root mean square residual correlations should be on the order of 0.05 or less and there should not be many individual residual correlations above 0.1.

The methods described above constitute one approach, albeit one that we consider to be the best generally available using current psychometric methods. There are obviously other methods that could also be employed in conjunction with or in place of those described above. A full discussion of the relative merits of alternative approaches would be useful, but is beyond the scope of this paper.

DETERMINING THE ESSENTIAL UNIDIMENSIONAL COMPONENTS OF THE MKT TEST DOMAIN: AN EMPIRICAL INVESTIGATION

Data Source

Items were piloted in California's Mathematics Professional Development Institutes. These institutes were publicly funded, large-scale efforts to boost California teachers' knowledge of subject matter in mathematics. Piloting of two forms took place with elementary teachers. Form A consisted of 12, 14, and 20 CCK, SCK, and KCS items respectively while for form B, the distribution was 12, 12, and 19 for CCK, SCK, and KCS. These two forms had 12 items, or "linking items," in common. We collected 640 cases for form A and 595 for form B. All these items involved number concepts and operations.

Item format was most often single multiple choice with four or more alternatives as in question 2 of the appendix. However, a number of questions had the form given in question 1 of the appendix, where a single scenario or stem had three separate sub-questions each with two alternatives (e.g., "yes"/"no") and a choice of "I'm not sure." Because the sub-questions were all related to a single stem, these items were considered to be "testlets" (Wainer & Keily, 1987). Testlets were scored by summing the number of correct response to sub-questions within a stem. The resulting data was then analyzed using a multidimensional generalization of Samejima's (1969) unidimensional IRT model for rating scale data implemented in the computer program ORDFAC (Schilling, 2005).

Theoretical Specification of the MKT Domain in Terms of Essentially Unidimensional Components

The first step in assessing our interpretive argument is to use theory underlying the measures to specify conceptual unidimensional subdomains. In the case of the MKT measures, the substantive theory states that content knowledge for teaching mathematics includes both basic mathematics *and* knowledge that is specific to the work of teaching. This translates into seven categories of knowledge: 1) knowledge of the mathematical content taught in elementary schools; 2) providing explanations for mathematical ideas and procedures; 3) representing ideas and procedures using number lines, area models, or word problems; 4) determining the correctness of alternative or non-standard mathematical methods; 5) understanding typical student errors; 6) assessing the degree to which student responses to questions indicate understanding of mathematical concepts; and 7) ordering problems in terms of student difficulty. Based on this specification, and as stated in Assumption 2 and inferences 2A and 2B, the MKT domain can be conceptually subdivided on the basis of types of knowledge. Content knowledge (CK) can be considered its own construct (categories 1–4) or as the sum of two sub-components, common content knowledge (CCK)—the mathematical knowledge that is widely available and shared between professions (category 1), and specialized content knowledge (SCK)— teaching-specific mathematics (categories 2 through 4). Knowledge of content and students (KCS) includes teaching-specific knowledge focused on student understanding (categories 5 through 7).

ASSESSING ESSENTIAL UNIDIMENSIONALITY OF THE MKT SCALES—RESULTS

Our first step in decomposing the MKT items in terms of essentially unidimensional components consisted of specifying a theoretical structure in terms of CCK, SCK, and KCS. Our second step consists of assessing the dimensionality of the items by fitting full-information item factor analysis (FIIF) models of varying dimensions and assessing their fit by using likelihood ratio chi-square statistics. These statistics, along with the LR p-values and the AIC indices, are presented in Table 1.

The AIC criterion indicates a two-factor model for form A and a three-factor model for form B. However, an examination of the three-factor solution for form B failed to reveal an interpretable third factor. Therefore in Table 2 we present the Varimax (Kaiser, 1958) rotated factor loadings for two-factor models for both forms A and B, with loadings greater than 0.3 highlighted in boldface. A cursory examination of the factor loadings reveals that the CCK and SCK items load primarily on the first factor (i.e., a CK factor) and the KCS item load primarily on the second factor, although there are some exceptions. The question now is whether this decomposition into two factors yields essentially unidimensional components or whether further decomposition or fine tuning is necessary.

In order to explore this in more detail we present item vector plots of the CK and KCS items for forms A and B in Figures 1 and 2. Item difficulties with respect to the first factor are indicated on the horizontal axis moving from left to right; item difficulties with respect to the second factor are indicated on the vertical axis from bottom to top. The first panels of these figures present the CK and KCS item vectors together; the second panels present the KCS items separately in order to determine if the KCS items can constitute an essentially unidimensional scale. Later analyses will examine the CK items separately and in more detail.

TABLE 1
Fit of FIIF Analyses of CK/KCS Items—Form A and B

Form	*Model*	*Chi-square*	*df*	*p*	*AIC*
A	1 Factor				25253
	2 Factor	97	26	0.000	25208
	3 Factor	45	25	0.009	25213
	4 Factor	47	24	0.003	25214
B	1 Factor				22422
	2 Factor	73	29	0.000	22407
	3 Factor	66	28	0.000	22397
	4 Factor	49	27	0.006	22402

TABLE 2
Varimax Rotated Factor Loadings: Two Factor Model

Form A				*Form B*			
Item	*Type*	*Factor 1*	*Factor 2*	*Item*	*Type*	*Factor 1*	*Factor 2*
LCCK1	CCK	**0.53**	**0.39**	lnc1	CCK	**0.44**	**0.52**
LCCK2	CCK	**0.43**	0.26	lnc2	CCK	**0.32**	**0.36**
LCCK3	CCK	**0.30**	0.03	lnc3	CCK	0.26	0.26
A1	CCK	0.12	0.07	B1T	SCK	**0.39**	0.10
A2T	SCK	**0.36**	0.28	B2	SCK	**0.42**	0.27
A5T	CCK	**0.28**	0.22	B4	CCK	**0.68**	**0.41**
A6	SCK	**0.30**	0.14	B5	CCK	**0.50**	**0.51**
LSCK1	SCK	**0.56**	0.29	B7	SCK	0.09	**0.44**
LSCK2	SCK	**0.33**	0.23	B11	CCK	**0.60**	0.23
LSCK3	SCK	**0.62**	0.08	B12	CCK	**0.54**	**0.35**
A15T	CCK	**0.62**	0.32	lop1	SCK	**0.78**	0.15
A18T	SCK	**0.55**	0.01	lop2	SCK	**0.42**	0.20
A19T	SCK	**0.41**	0.04	lop3	SCK	**0.47**	0.02
LKCS1	KCS	0.22	**0.32**	B18T	CCK	**0.44**	**0.45**
LKCS2	KCS	**0.42**	**0.48**	B20T	SCK	**0.40**	0.30
LKCS3	KCS	**0.30**	0.26	LNI1	KCS	0.14	0.21
A8	KCS	0.11	**0.35**	LNI2	KCS	**0.38**	**0.49**
A9	KCS	0.10	0.27	LNI3	KCS	0.14	**0.38**
A10	KCS	0.29	**0.47**	B9	KCS	**0.41**	**0.33**
A14T	KCS	**0.31**	0.11	B13T	KCS	0.35	**0.44**
LKCS4	KCS	0.11	**0.89**	LOI1	KCS	0.08	**0.37**
LKCS5	KCS	**0.43**	**0.35**	LOI2	KCS	0.27	**0.45**
LKCS6	KCS	0.16	**0.30**	LOI3	KCS	−0.10	**0.54**
A22T	KCS	**0.45**	0.20	B21	KCS	0.28	0.05
A23	KCS	0.00	**0.56**	B22	KCS	0.18	0.17
A25	KCS	0.07	**0.40**	B26	KCS	0.25	**0.44**
A27	KCS	0.10	0.20	B27T	KCS	0.26	**0.33**
				B28	KCS	0.16	0.22
				B29	KCS	0.19	**0.49**

The first panels of both plots reveal the separation of items into the two groups that we observed in Table 2. The CK items generally span a smaller vector in two-dimensional space, while the KCS items vary widely in their orientation. This also can be seen in the first two columns of Table 3, which present the item sector widths for the CK and KCS items. The item sector widths for the form A and B KCS items are 70 and 91 degrees, while those for the CK items are 39 and 75 degrees, respectively. However, the latter value for form B is somewhat misleading since it is almost entirely due to a single item—B7 which loads almost exclusively on the second factor in Table 2. Absent this item, the item sector width reduces to 46 degrees, which is in line with that observed for form A.

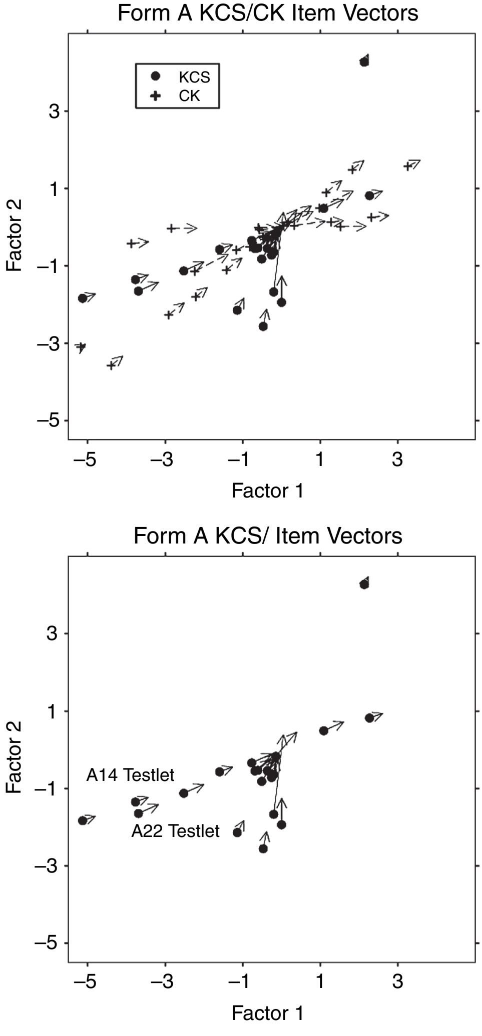

FIGURE 1 KCS/CK Plots—Form A.

Table 3 also shows that neither the CK nor the KCS items appear to comprise essentially unidimensional scales without some further revision. The KCS items in particular have large RMS residual correlations and large AIC GFI's, where an AIC GFI greater than one indicates that a set of items are not unidimensional. The CK items are better behaved with respect to the RMS residual correlations, but even here the goodness of fit indices indicate two dimensions. We will

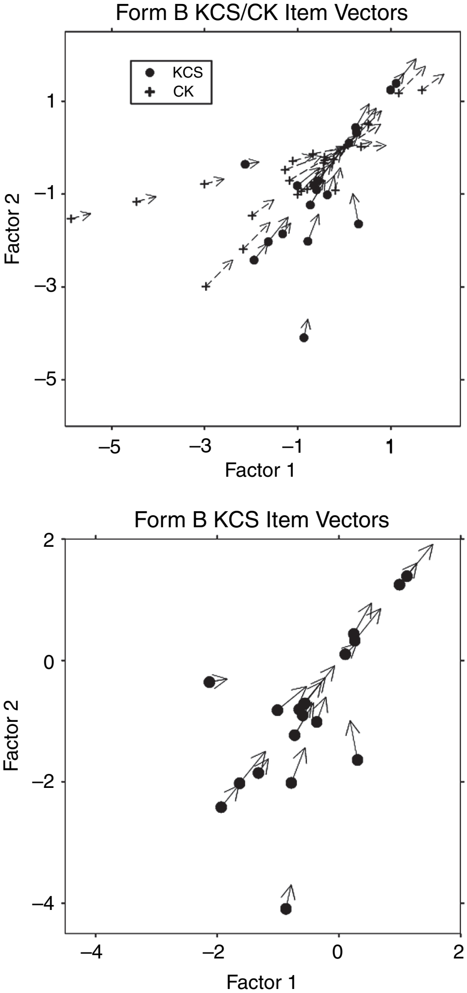

FIGURE 2 KCS/CK Plots—Form B.

subsequently explore the dimensionality of the CK items with respect to the CCK/SCK distinction, but first it is instructive to examine the plots of the KCS item vectors separately in the second panels of Figures 1 and 2.

These plots indicate that most of the items are oriented in a similar direction at about 60 degrees relative to the horizontal axis; the ratio of the loadings on second factor relative to the first factor is therefore about two to one. However, there are a few items in both forms that load more heavily on the first factor—two testlets and

TABLE 3
Fit of Essential Unidimensionality: CK and KCS Scales

	Item Sector Width		*AIC GFI*		*RMS Residuals*	
Scale	*Form A*	*Form B*	*Form A*	*Form B*	*Form A*	*Form B*
CK	38.5	75.4	1.26	1.28	0.059	0.057
KCS	70.2	91.0	1.40	1.14	0.073	0.068
Revised KCS	48.6	26.9	1.11	0.27	0.079	0.058

a single item on form A and a single item on form B. Form B also has a single item which loads almost exclusively on the second factor, an item dealing with student errors in addition. The items in form A are lacking with respect to moderate to difficult items, with only a single difficult item, while the items of form B are better in this regard. If we revise the KCS scales by excluding the items that are outliers in terms of their orientation, we are able to obtain a good essentially unidimensional scale for the form B KCS items, with a small item sector width, AIC GFI, and RMS residual correlation. However, revision of the form A KCS scale does not have as great an effect—the item sector width, AIC GFI, and RMS residual correlation indicate that this scale is still two dimensional.

Further analysis of the CK items for both forms revealed that a two-dimensional model was needed for both forms, as indicated in Table 4. Plots of the item vectors for both forms are given in Figure 3; fit statistics for essential unidimensionality are given in Table 5. With the exception of a single item for form A that suffered from poor wording, all of the CCK items are tightly clustered in a narrow sector for both forms; the item sector widths are approximately 30 degrees or less with small AIC GFI's and RMS residual correlation. However, the SCK items vary widely in their orientation, with large validity sector widths, AIC GFI's and RMS residual correlations. The exception to this trend is the AIC GFI for form B SCK, but this low value (0.84) occurs because only two items load on the second factor.

TABLE 4
Fit of FIIF Analyses of CCK/SCK Items: Form A and B

Form	*Model*	*Chi-square*	*n.p.*	*p*	*AIC*
A	1 Factor		13		14129
	2 Factor	30	25	0.003	14123
	3 Factor	13	36	0.313	14132
B	1 Factor		15		11911
	2 Factor	36	29	0.001	11903
	3 Factor	18	42	0.171	11911

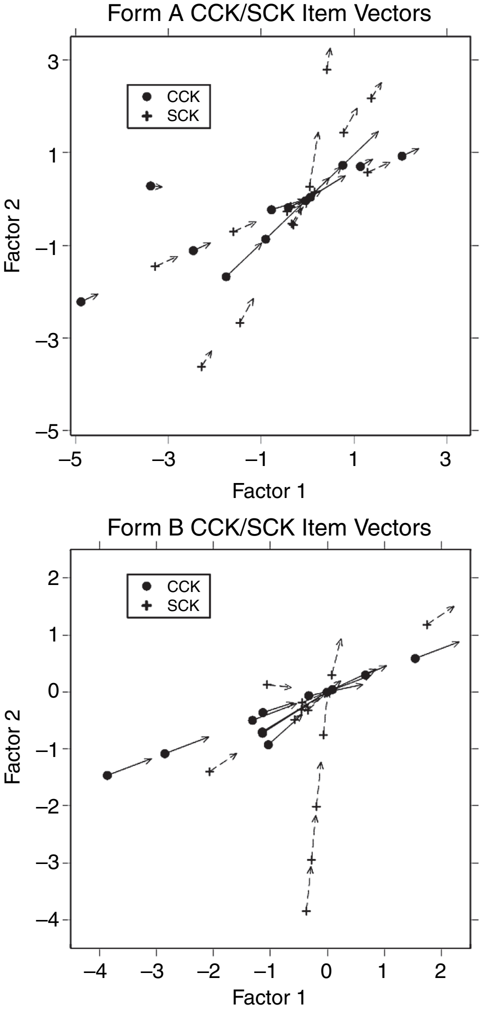

FIGURE 3 CCK/SCK Plots—Forms A and B.

However, many of the SCK items are oriented in the same direction as the CCK items, suggesting that revising the overall scale to include only those SCK items could produce an essentially unidimensional scale. The CCK items and selected SCK items are considered together in the third row of Table 5 and present an improvement with regards to essential unidimensionality, with smaller item sector widths, and RMS residual correlations than was previously observed. Moreover, the AIC GFI's indicate that this collection of items is essentially unidimensional.

TABLE 5
Fit of Essential Unidimensionality: CCK and SCK Scales

	Item sector width		*AIC GFI*		*RMS Residuals*	
Scale	*Form A*	*Form B*	*Form A*	*Form B*	*Form A*	*Form B*
CCK	27.0	30.3	0.84	0.36	0.051	0.045
SCK	57.6	77.7	1.54	0.84	0.059	0.053
Revised CK	33.1	38.1	0.90	0.64	0.055	0.047

CONCLUSIONS

What conclusion can we draw concerning the MKT measures and about the role of psychometric analysis within the validity argument approach?

Our first inference that the organization of mathematical knowledge for teaching into common content knowledge, specialized content knowledge, and knowledge of content and students would result in multiple factors in the item factor analyses of the scales was clearly supported. The FIIF analyses clearly showed that the measures were multidimensional and that the constructs of KCS and SCK were responsible for the multidimensionality.

However, with respect to the second inference, that we could construct essentially unidimensional scales reflecting the CCK/SCK/KCS organization that would reliably distinguish teachers, the results were decidedly more mixed. By any index, CCK was essentially unidimensional and the CCK items could be used to construct reliable and essentially unidimensional scales.

KCS items, as the name suggests, involves two components—knowledge of students and knowledge of content, with most KCS items reflecting these components at a 2/1 ratio. As Reckase (1988) suggests, unidimensional scales can be constructed from multidimensional items if the ratio of the components can remain relatively homogeneous. In fact, we were able to do this for form B—a scale with 10 out of the 14 KCS items was essentially unidimensional and achieved a reliability of 0.67, which is acceptable considering the small number of items.

But care needs to be taken to achieve essential unidimensionality. For example, we were not able to achieve an essentially unidimensional and reliable KCS scale using the form A items, partly because the validity sector of the items remaining after excluding extremely-oriented items was still large—49 degrees. Narrowing the validity sector for form A produces an essentially unidimensional set of items, but a set too few in number and too homogeneous in difficulty to produce reliable measurement. An alternative to constructing essentially unidimensional scales out of items that are inherently multidimensional is to use multidimensional IRT to simultaneously measure both factors. Unfortunately, our sample sizes do not permit such an approach.

With respect to SCK there is probably no way to construct an essentially unidimensional scale. SCK items, while clearly introducing multidimensionality into content knowledge for teaching, vary even more widely in their orientation than the KCS items. Instead of thinking of SCK as a different type of knowledge or a separate dimension, it is more instructive to think of SCK as specialized knowledge about mathematics that becomes available to teachers in differing degrees due to varying factors in the environment, such as professional development efforts for teaching mathematics or changing concepts of mathematical knowledge for teaching in teacher education programs. This would account for the inconsistent nature of the results for SCK – in fact, subsequent pilot data have revealed that sometimes SCK shows up as a separate factor in factor analyses and sometimes it does not.

With respect to the role of psychometric analysis in test validation our example also allows us to draw some important conclusions. The central idea behind Kane's validity argument approach is that test scores are "interpreted" in order to make informed decisions; therefore possible interpretations need to clearly stated, and evidence needs to be provided for the interpretations. This is what test validation is all about. The more detailed the interpretations and evidence, the better the understanding of the proper interpretation of a test score and the limits on that interpretation. Clearly the specification of a test's structure in terms of essentially unidimensional components and checking those assumptions via MIRT provides a more detailed level of interpretation and evidence than the amorphous concept of a test domain from which items are randomly sampled and simple reliability or generalizability indices are obtained. Essential unidimensionality is one of the key concepts of modern psychometric theory and it behooves any test validation effort to investigate implicit or explicit assumptions of essential unidimensionality underlying a particular measure.

This is not meant to criticize generalizability theory, because even generalizability theory is best employed when the test domain is structured, such as with a multi-faceted test domain where variance components measure the importance of the different facets. Rather our goal is to focus attention on the importance of structure no matter what school of psychometrics is employed.

Test validation should not be a checklist-driven yes-or-no process. In our case, the exploration of the structure of the MKT domain provided us with a better understanding of the concept of mathematical knowledge for teaching, our conception of the test domain, and the limits and boundaries on the use of test scores arising out of our measures. Such an enhanced understanding is the true measure of a test validation process.