A validation study of the use of mathematical knowledge for teaching measures in Ireland

Seán Delaney

Coláiste Mhuire, Marino Institute of Education, Dublin, Ireland

Tel: + 353 1 805 7722

Fax: + 353 833 5290

E-mail: sean.delaney@mie.ie

Abstract

Researchers who study mathematical knowledge for teaching (MKT) are interested in how teachers deploy their mathematical knowledge in the classroom to enhance instruction and student learning. However, little data exists on how teachers' scores on the US-developed measures relate to classroom instruction in other countries. This article documents a validation study of Irish teachers' scores on measures of MKT that were adapted for use in Ireland. A validity argument is made identifying elemental, structural and ecological assumptions. The argument is evaluated using qualitative and quantitative data to analyse inferences related to the three assumptions. The data confirmed the elemental assumption but confirming the structural and ecological assumptions was more difficult. Only a weak association was found between teachers' MKT scores and the mathematical quality of instruction. Possible reasons for this are outlined and challenges in validating the use of measures are identified.

Keywords

Mathematical knowledge for teaching, measures, mathematical quality of instruction, validity, validation study, cross-cultural, elementary school

Abbreviations

3-D	Three-dimensional
CCK	Common content knowledge
CK	Content knowledge
COACTIV	Cognitive Activation in the Classroom
IRT	Item response theory
KCS	Knowledge of content and students
KCT	Knowledge of content and teaching
MKT	Mathematical knowledge for teaching
MQI	Mathematical quality of instruction
OECD	Organisation for Economic Co-operation and Development
PISA	Programme for International Student Assessment
SCK	Specialized content knowledge
TEDS-M	Teacher Education Study in Mathematics
TIMSS	Trends in International Mathematics and Science Study
US	United States

A Validation Study of the Use of Adapted Mathematical Knowledge for Teaching Measures in Ireland

1. Background

1.1 Overview

Over the last few decades researchers have become more interested in teachers' knowledge generally. Inspired by the work of Shulman (e.g. 1986, 1987), this interest has been prompted by greater conceptual understanding of the knowledge teachers hold and use. Several empirical studies have looked at the knowledge held by teachers and by student teachers (e.g. An, Kulm, & Wu, 2004; Ma, 1999; Tatto et al., 2008). In addition, some policy reviews, including reviews conducted by the Organisation for Economic Co-Operation and Development (OECD) (e.g. 2004, 2008), have identified shortcomings in teachers' subject matter knowledge. Teacher knowledge of mathematics has been a particular focus of this attention.

Given the widespread interest in teachers' mathematical knowledge, it is not surprising that an instrument that measures their mathematical knowledge at scale (Hill, Schilling, & Ball, 2004) would be of interest to scholars around the world. To date the measures developed by Ball, Hill and colleagues in the United States have been used in several countries, including Ghana (Cole, this issue), Indonesia (Ng, this issue), Ireland, Korea (Kwon & Pang, this issue), and Norway (Fauskanger et al, this issue). Such widespread use of the measures should not be unexpected because adapting educational tests for use in additional languages and cultures is cheaper and faster than developing a new one (Hambleton & Kanjee, 1995).

However, when a test is used in any setting, evidence must be presented to justify the proposed interpretation and use of the test results (Messick, 1989). This is done to establish test validity and is particularly important when the test is to be used in a setting outside that for which it was designed (van de Vijver & Leung, 1997). Guidelines have been developed for establishing the validity of international tests in mathematics, science and literacy, particularly the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) (Hambleton & Kanjee, 1995; OECD, 2009). However, these guidelines cannot automatically be applied to a test of teacher knowledge for at least four reasons. First, most tests of teacher knowledge tend to be small scale and do not have the resources available to large testing endeavours such as TIMSS and PISA. Second, when researchers use US measures of teacher knowledge, the measures have already been developed and used in the United States whereas in TIMSS and PISA, measures are developed and amended across countries prior to being used in any one country. Third, TIMSS and PISA tests are deliberately designed to compare mathematical performance across countries and thus the test results are likely to be used in a similar way across countries. The results of tests of teacher knowledge may be used for different purposes in different countries (e.g. to inform initial teacher education, to evaluate outcomes of continuous professional development programmes, to examine a relationship between teacher knowledge and student learning) with different implications for validation. Finally, although mathematical knowledge is widely assumed to be universal, the knowledge required for teaching may be culturally based (Andrews, 2011; Stylianides & Delaney, 2011). Indeed, some researchers have noted the value of comparing conceptualizations of teaching expertise - including but not limited to knowledge - across countries (Li & Even, 2011; Yang & Leung, 2011).

This paper describes a validation study designed to determine if measures of mathematical knowledge for teaching (MKT) developed in the United States, and adapted for use in Ireland, can be validly used to study the mathematical knowledge of Irish teachers. The paper begins with an overview of the theory and construct of mathematical knowledge for teaching before outlining conceptions of validity. A description of the research design is presented next. This is followed by the results and a discussion of the results.

1.2 Mathematical knowledge for teaching

Shulman proposed three categories of content knowledge that are important for teachers: subject matter content knowledge, pedagogical content knowledge, and curricular knowledge (Shulman, 1986). Shulman’s work has been the stimulus for several researchers who have applied his theory to various school subjects, including mathematics. One of the few studies that has attempted to study teacher knowledge across countries is the Teacher Education Study in Mathematics (TEDS-M). Schmidt and his colleagues (2008) studied the opportunities to learn that prospective teachers had in 21 countries. They consider professional competence to consist of professional knowledge and professional beliefs (Schmidt et al., 2007). Professional knowledge consists of content knowledge and pedagogical content knowledge, which in turn is made up of instructional planning, student learning and curricular knowledge. Professional beliefs are made up of epistemological beliefs regarding mathematics, instructionally related beliefs about teaching and instructionally related beliefs about how students learn mathematics. However most research on teacher knowledge, some of which has informed the work of the TEDS-M study, is based in individual countries.

In England, Rowland and his colleagues used Shulman's categories of subject matter knowledge and pedagogical content knowledge to develop the knowledge quartet (Rowland, Huckstep, & Thwaites, 2005), a framework focused on the situations in which teacher knowledge comes into play (Turner & Rowland, 2011). Rather than being specifically mapped to one or other of Shulman's categories of subject matter knowledge and pedagogical content knowledge, each dimension of the knowledge quartet (Rowland et al., 2005) relates to a combination of Shulman's two categories.

In contrast, the German COACTIV (Cognitive Activation in the Classroom) research group used Shulman's categories to conceptualize the subject matter knowledge that secondary teachers need "to be well prepared for their instructional tasks" (Baumert et al., 2010, p. 141). They hypothesise three subscales within pedagogical content knowledge: knowledge of explanations and representations, knowledge of students' thinking, and knowledge of multiple solutions to mathematical tasks (Krauss, Baumert, & Blum, 2008). Subject matter knowledge, or content knowledge, is conceptualized as "deep understanding of the contents of the secondary school mathematics curriculum" (p. 876). Although Krauss and his colleagues (2008) found pedagogical content knowledge (PCK) and content knowledge (CK) to be related, their validation study found that PCK and CK constitute different dimensions of teachers' professional knowledge.

Ball, Thames and Phelps (2008) at the University of Michigan have proposed another refinement of Shulman's categories, with specific reference to mathematics. Their theory, MKT, was developed by analysing the work of teaching from a mathematical perspective (Ball & Bass, 2003). Their elaboration of pedagogical content knowledge bears some resemblance to the student and instruction subscales of Krauss et al (2008). Ball and colleagues (2008) include knowledge of content and students (KCS, knowledge of content as it relates to students) and knowledge of content and teaching (KCT, knowledge of mathematics and its relationship to the tasks of teaching). Shulman's curricular knowledge is also included as part of pedagogical content knowledge in this model. The Michigan conception of subject matter knowledge is broader than that of COACTIV and it includes common content knowledge (CCK), a category that is deliberately omitted by the COACTIV group. Under subject matter knowledge, Ball, Thames and Phelps (2008) include CCK (the mathematical knowledge used by people generally in their life and work), specialized content knowledge (SCK, specific knowledge needed by teachers to do the mathematical work of teaching) and a provisional category of horizon content knowledge (knowledge of the connections among mathematical topics on the curriculum). Even if these six categories of knowledge are not definitive, content knowledge for teaching is likely to be multidimensional (Ball et al., 2008).

Both the Michigan group and the German group have developed measures to study teachers' mathematical knowledge. The Michigan group originally focused on studying primary teachers using multiple-choice measures.[1] The German group's measures are open-ended and focused on secondary teachers' mathematical knowledge. The measures, which are the result of substantial investment of resources, may be attractive to researchers outside of the United States and Germany. Indeed, some researchers have described MKT as "the most promising current answer to the longstanding question of what kind of content knowledge is needed to teach mathematics well" (Morris, Hiebert, & Spitzer, 2009, p. 492). Yet, there is a need to be cautious in using measures developed in one country in another country, especially when these measures are explicitly grounded in the practice of teaching and the practice of teaching can vary across countries (Stigler & Hiebert, 1999).

Andrews (2011) is critical of existing frameworks which view teacher knowledge as located within the individual without reference to the cultural context in which the teacher works. In response to this problem Andrews proposes using overarching frameworks in light of the idealized, received and intended curricula of an individual teacher or a of a country. Pepin (2011) provides a concrete example of how one task of teaching - listening - can vary across countries. Listening may be practised with attention to individual pupils (England) or with attention to the whole class (France). Even within a country (Germany), the kind of listening practised could focus either on pastoral support of students among Hauptschule teachers or on mathematics among Gymnasium teachers. Such variations provide reason to be cautious when applying constructs in countries other than the country in which they were developed. Therefore when MKT measures were adapted and used to study Irish teachers' mathematical knowledge, a validation study was required.

1.3 Establishing Validity of Adapted Measures of MKT

Test validation is concerned with demonstrating that a teacher's score on MKT measures is an indication of the teacher's mathematical knowledge for teaching and not an indication of some other factor (Messick, 1989), such as the teacher's general mathematical knowledge, the teacher's general intelligence or the teacher's test-taking skills. Establishing validity takes on a particular importance if the test results have consequences for the teacher, such as certification, promotion or tenure. However, validity in educational research is a contested issue (Lissitz & Samuelsen, 2007) and was the subject of a dialogue in a 2007 issue of *Educational Researcher* (volume 36, number 8). In addition to the contested nature of validity, its implementation is often disconnected from its conceptualization (Schilling & Hill, 2007). Furthermore, in a study such as the present one, where measures based on a US construct are used in Ireland, cultural factors mean that there is a risk of cultural bias in using the test of MKT and this must be considered as part of the validation study (Hitchcock et al., 2005).

Three categories of test validation have typically been used: criterion validity, content validity and construct validity. With criterion validity a test result is compared with a stated criterion (e.g. performance in first year of college) but finding a criterion against which to compare the results can sometimes be difficult (Kane, 2006). Content validity, is established not with reference to a particular criterion but by claiming that performance on a sample of tasks from a domain estimates one's overall performance in the domain, such as academic achievement. This form of validity is important but limited to interpreting scores "in terms of expected performance over some universe of possible performances" (Kane, 2006, p. 19) . A third type of validity, construct validity, began as a means of assessing the extent to which a test was an "adequate measure" (p. 20) of a particular theory .

In advocating a unified view of validity, Messick (1989) claimed that "because content- and criterion-related evidence contribute to score meaning, they have come to be recognized as aspects of construct validity" (p. 9), which implied that construct-validity was the only category that needed to be considered. Although this view has been contested, no consensus appears to have emerged around a new validity paradigm as evident in comments responding to Lissitz and Samuelsen (2007) (e.g. Sireci, 2007). In advocating an argument-based approach to validation, Kane (2008) characterizes the debate around validity as being "whether to retain a broad conception of validity as an evaluation of the proposed interpretation and uses of test results (e.g. Cronbach, 1971; Messick, 1989) or to adopt a narrower, operational definition of validity as an evaluation of the internal characteristics of the test" (Kane, 2008, pp. 76-77). The validation study described here represents an application of Kane's approach.

Validation in Kane’s model requires two steps. One is to propose an interpretive argument stating how results will be interpreted and used “by laying out the network of inferences and assumptions leading from observed performances to conclusions and decisions based on the performances” (Kane, 2006, p. 23). In the second step the plausibility of the proposed interpretive argument is evaluated. To illustrate how this works in practice, Kane (2004) applied his model to a specific case where test results are used to certify someone to practice in an area such as teaching. According to Kane (2004) the following steps require validation: (a) from participants’ observed performance on test items to a specific score; (b) from the specific score to a generalized score over all the test domain; (c) from the test domain to the required knowledge, skills and judgment domain; (d) from the knowledge, skills and judgment domain to the practice domain; (e) from the practice domain to certification.

Many assumptions and inferences are made in moving through these steps from performance on a test to being certified as fit for a field of practice. The inferences and assumptions in each step are different and are validated differently. Performance on the MKT measures cannot be used as a criterion for certification, hiring or promotion of teachers; therefore, specification of the steps (a) to (d) above is relevant to how scores on MKT items are interpreted and used. The test results are not used as an end in themselves but as a means to better understand knowledge for mathematics instruction. In a series of papers published in *Measurement: Interdisciplinary Research and Perspective* (Vol. 5, 2 and 3), Hill, Schilling and colleagues have applied Kane’s approach to the interpretation and use of MKT measures in the United States.

Schilling and Hill (2007) renamed steps (a), (b), (c) and (d) above and related them to three sets of assumptions and related inferences: elemental, structural and ecological. The elemental assumption [step (a)] relates to individual items in a test and how well the items capture teachers’ MKT, and not irrelevant factors such as test-taking strategies. The second assumption tested by Schilling and Hill, structural assumptions and inferences, relates to whether the MKT scales (or subscales) measure no more and no less than the domain of MKT (or its sub-domains CCK, SCK, KCS). This assumption incorporates Kane’s second and third inference because it concerns the extent to which a teacher's observed test score relates to the teacher's overall expected score on MKT and to the specific sub-domains of MKT. Schilling and Hill’s third category [step (d)] relates to ecological assumptions and inferences. This step validates teachers’ levels of MKT in light of how their MKT affects their practice. The assumption is that adapted MKT measures capture teacher knowledge that is related to effective mathematics instruction. In this paper all three assumptions – elemental, structural and ecological – relating to the use of the MKT measures in Ireland will be investigated.

In a previous article (Delaney, Ball, Hill, Schilling, & Zopf, 2008), reasons for adapting US MKT measures for use in Ireland were outlined. Four categories of changes were identified and applied to a set of MKT measures. Adaptations were made to the items' general cultural context, their school cultural context, the mathematical substance and other changes.

Based on the adaptations made and on the validity imperative, the research questions for this article are: (1) Can the adapted MKT measures be validly used to make claims about the MKT held by a large group of teachers in Ireland? (2) What are the challenges in conducting a validation study of adapted MKT measures across countries? Although Hill and her colleagues have validated assumptions of MKT for use in the United States, separate validation is required for Ireland in order to investigate the validity of using adapted measures in a new setting.

2. Method

The first step in validation is to make an interpretive argument. According to Kane the interpretive argument "specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from observed performances to conclusions and decisions based on the performances" (Kane, 2006, p. 26). This argument is then evaluated for its coherence, for the reasonableness of its inferences, and for the plausibility of its assumptions (Kane, 2006). The full interpretive argument for using the MKT measures in Ireland is as follows (adapted from Schilling & Hill, 2007):

(1) Elemental assumption: Teachers used their MKT when responding to questions on the form.

Inference: A teacher's chosen response to a particular item was consistent with their mathematical reasoning about the item.

(2) Structural assumption: The domain of mathematical knowledge for teaching can be distinguished by the types of knowledge deployed by teachers (i.e. CCK, SCK and KCS).

Inference: Items will reflect this organization with respect to the type of knowledge held.

(3) Ecological assumption: The MKT multiple-choice measures captured the mathematical knowledge teachers need to teach mathematics effectively.

Inference: Teachers' scale scores on the measures are related to the quality of the teachers' mathematics instruction. Higher MKT scale scores are related to more effective mathematics instruction and lower scale scores are related to less effective mathematics instruction.

This study is part of a larger study in which three different sets of teachers were recruited and studied. First, a convenience sample of 100 primary teachers was recruited to pilot MKT items that had been adapted for use in Ireland and five of these teachers were interviewed about their answers to the pilot items (see Delaney et al., 2008). Second, a national sample of 501 Irish teachers completed a test to measure their MKT (see Delaney, 2008). Third, ten additional Irish primary teachers completed a test of the MKT measures and were videotaped teaching four mathematics lessons each.[2] Most of the ten teachers were recruited by asking teacher educator and school principal acquaintances to recommend teachers who might be willing to be videotaped teaching mathematics, and two were recommended by teachers who had already been videotaped. The goal was to recruit typical teachers to study "typical case" (Patton, 2002) samples of Irish mathematics teaching; but possibly teachers with lower levels of mathematical knowledge were less likely to be recommended or to agree to be videotaped.

2.1 The elemental assumption

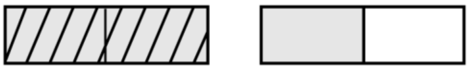
The purpose of evaluating the elemental assumption is to ascertain whether teachers responding to the items used their MKT, or their general knowledge of teaching, or test-taking strategies. Following the pilot study, five teachers, who had responded to the set of adapted items, were interviewed. The teachers were chosen based on their willingness to spend one extra hour answering questions about the test. Twelve adapted items[3] were selected (including two testlet items, one of which had four parts attached to one stem and another that had three parts attached to one stem) to include representative questions on SCK (number and operations, algebra, and geometry; see Figure 1) and KCS (see Figure 2). Interviewees were asked why they gave the answer they gave; the interviews were recorded and subsequently transcribed. Responses were analysed and coded for being consistent or inconsistent with the written response to the multiple-choice questions (Hill, Dean, & Goffney, 2007). The thinking used by the five respondents to answer the twelve items was also analysed and categorized using codes developed and described by Hill, Dean and Goffney (2007). The codes are mathematical justification, memorized rules or algorithm, definitions, examples/counterexamples, other mathematical reasoning (all drawing on mathematical knowledge); knowledge of students and content (drawing on knowledge of students); and guessing, test-taking skills, and other non-mathematical thinking.

In order to support the elemental inference, it would be expected that the teachers' thinking about responses would be consistent with their chosen answers. In other words, if the teacher answered correctly, their reasoning should support that answer and if a teacher responded incorrectly, they should demonstrate lack of understanding of the topic. It would also be expected that the teachers used their knowledge of mathematics or of students in responding to the items and not generic test-taking skills or guessing.

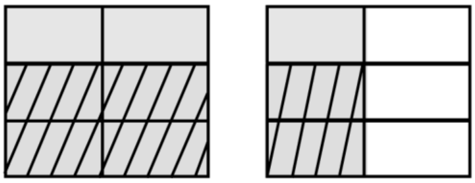
At a professional development workshop, teachers were learning about different ways to represent multiplication of fractions problems. The leader also helped them to become aware of examples that do not represent multiplication of fractions appropriately.

Which model below cannot be used to show that $1\frac{1}{2} \text{ x } \frac{2}{3} = 1$? (Mark ONE answer.)

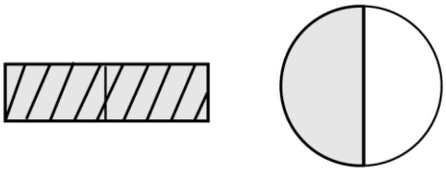
A)



B)



C)



D)

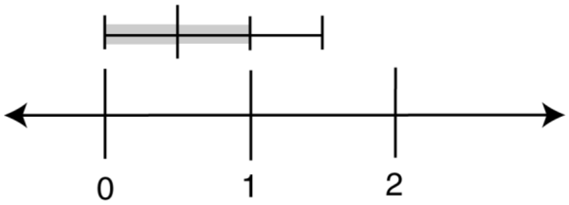


Figure 1. Sample SCK item (no adaptation necessary). Taken from http://sitemaker.umich.edu/lmt/files/LMT_sample_items.pdf.

Mrs. McKenna is planning mini-lessons for students focused on particular difficulties that they are having with adding columns of numbers. To target her instruction more effectively, she wants to work with groups of students who are making the same kind of error, so she looks at some recent classwork to see what they tend to do. She sees the following three student mistakes:

$$\text{I)}\quad \begin{array}{r} {\scriptstyle 1} \\ 38 \\ 49 \\ +\ 65 \\ \hline 142 \end{array} \qquad\qquad \text{II)}\quad \begin{array}{r} {\scriptstyle 1} \\ 45 \\ 37 \\ +\ 29 \\ \hline 101 \end{array} \qquad\qquad \text{III)}\quad \begin{array}{r} {\scriptstyle 1} \\ 32 \\ 14 \\ +\ 19 \\ \hline 64 \end{array}$$

Which have the same kind of error? (Mark ONE answer.)

a) I and II

b) I and III

c) II and III

d) I, II, and III

Figure 2. Sample (adapted) KCS item. Original taken from http://sitemaker.umich.edu/lmt/files/LMT_sample_items.pdf.

2.2 The structural assumption

The second assumption in the validity argument is that the domain of MKT can be distinguished by the types of knowledge used by teachers in responding to the measures. A factor analysis was conducted on the responses of 501 teachers to the MKT test. The teachers were drawn from a national, random representative sample of Irish primary schools. Data from the teachers interviewed for the pilot study was also used to determine the extent to which teachers drew on SCK to answer SCK items and on KCS to answer KCS items.

In order to support the structural inference, it would ideally be expected that survey items that are considered conceptually to be SCK load on one factor, items considered to be CCK load on another and that KCS items load on a third. However, such a finding would be ambitious given that the items did not load so neatly in a similar study in the United States. One empirical US study found that SCK and CCK loaded on one factor, KCS on another and algebra on another (Hill et al., 2004). Such a finding need not be detrimental to the validity argument because it may be possible to modify or replace the inference and “remain consistent with both the assumption and related empirical evidence” (Schilling, Blunk, & Hill, 2007, p. 122).

2.3 The ecological assumption

The ecological assumption is concerned with the extent to which MKT multiple-choice measures captured the mathematical knowledge teachers need to teach mathematics effectively. This assumption is concerned with the relationship between teachers’ scores on the measures and the quality of the mathematics that they use when teaching. What is of interest here is not teachers' teaching strategies or style but the "mathematical content available to students during instruction" (Learning Mathematics for Teaching, 2011, p. 30). This has been conceptualised as the mathematical quality of instruction or MQI. In order to test this assumption, I used an instrument based on the MQI.

The MQI instrument that was used for this study consists of 32 features of mathematics instruction known as “codes” grouped in three sections,[4] and an accompanying glossary to explain the codes (Learning Mathematics for Teaching, 2006). The first group of codes reflects how teachers’ knowledge of the mathematical terrain of the enacted lesson is evident in instruction. Sample codes are the teacher’s use of technical language (e.g. equation, perimeter) and general language to describe a mathematical idea (e.g. referring to *exchanging* ten units for one ten); a teacher’s selection of representations and links made between or among them; and the presence of explanations.

The second category of codes refers to the teacher’s use of mathematics with students. Sample codes include how the teacher *uses* representations; how the teacher responds to students’ errors or expression of ideas; and whether the teacher elicits explanations from the students. The third category of codes considers the teacher’s use of mathematics to teach equitably in relation to inclusion and participation of students of all races and social classes. Sample codes include the amount of instructional time spent on mathematics; and the teacher’s encouragement of a diverse array of mathematical competence. In addition, coders gave a “global lesson score” to rate the teacher’s overall level of mathematical knowledge as low, medium or high on the basis of the instruction observed in the lesson. Given the range of codes to be considered, the process of coding needed to be consistent and explicit.

Lessons were divided into 5-min clips[5] for coding purposes (Learning Mathematics for Teaching, 2006). Two randomly paired members of the Learning Mathematics for Teaching team[6] were assigned to code a lesson. Team members who coded the lessons for this study came from Ghana (1), Ireland (1) and the United States (4). Each member watched the entire lesson, and then independently coded the MQI in each 5-min clip. Having independently coded the lessons, each coding pair met to reconcile their codes.[7] A narrative was written for each lesson noting salient points about its mathematical quality. In previous coding an inter-rater reliability check found that agreement among pairs ranged from 65% to 100% on individual codes (Learning Mathematics for Teaching, 2006).

I now return to the validity argument and its inferences. I wanted to evaluate the plausibility of the elemental, structural and ecological assumptions of the argument.

3. Results

3.1 The elemental assumption

The first stage in the validation process was to examine the extent to which teachers' written responses to questions were consistent with their thinking as articulated in follow-up interviews. For example, if a teacher's response to an item is wrong and the teacher does not have the knowledge being tapped, this is coded as being Wrong and Consistent (WC); whereas if a teacher gives a wrong response, despite having the relevant knowledge, this is coded as Wrong and Inconsistent (WI) because the response was inconsistent with the knowledge held. With five teachers and seventeen questions (including the testlet items), a total of 85 data points were possible (see Table 1). In almost three-quarters of the items (74%) the teachers' thinking was consistent with their written response. In 16.5% of the items it was not possible to determine if the teacher's thinking was consistent or inconsistent. This was usually because the interviewer did not press the respondent sufficiently for a response. In 9% of items, the teacher's thinking was inconsistent with the written response.

The reason for inconsistent responses varied. In one case a diagram used in the pilot questionnaire was found to be potentially ambiguous and this was corrected on that item in the final questionnaire. Another teacher's response was inconsistent because in a question that centred on the number of fractions between 0 and 1, he restricted his answer to the number of fractions that children would identify between 0 and 1. The dialogue begins after the teacher stated that there were not infinitely many fractions between 0 and 1.

I: When you say “Not infinitely many solutions”, amm, say, what kind of fractions would they come up with between 0 and 1?
R: Well, depending on, they’d come up with maybe two halves, three thirds, four quarters depending on how much experience in fractions they had, you know.
I: Right.
R: But they’d, from my experience anyway, in third and fourth, that’s what they’d say like. They’d just say the ones they kind of know and probably leave it at that, you know.
I: Ok, and if you took, say if you left the children out of it. Supposing, say it was an adult was answering that question, would they, would it also be “Not infinitely many solutions” for them?
R: Thinking of it now, it would be, ha, ha
I: It would be…?
R: Yeah
I: It would be this one [pointing to “infinitely many solutions”]
I: Ok, what might they come up with that the children wouldn’t come up with?
R: Well, they’d come up with sixteenths, twentieths, hundredths, thousandths, millionths, whatever.

The teacher justified his initial response by referring to his experience in third and fourth (grade). When he was prompted to consider a response that an adult would give, he changed the response and included denominators that they would use, that he thought would not be used by children. This is deemed to be inconsistent because the teacher had the relevant mathematical knowledge but by restricting the answer to what children might say, he had not recorded the correct answer in the written version of the test. There is little evidence to suggest that the inconsistent responses indicate problems with particular items because only in the case of one item did more than one teacher respond inconsistently. For this item, the reasons for being inconsistent differed: one teacher chose more than one relevant response rather than the “most important” one, which was required and the other could not adequately explain how she chose the correct answer.

Table 1

Analysis of consistent-inconsistent responses in post pilot test questionnaire

Item	Maria	Mark	Michael	Malcolm	Morgan
Specialized content knowledge items					
1A	RN	RC	RI	RN	RC
1B	RC	RC	RC	RC	RC
1C	RN	RC	RN	RC	RC
1D	WI	RC	RC	RC	RC
4A	RC	RC	WI	RC	RC
4B	RC	RI	WC	RN	RC
4C	RC	RC	RC	WI	RC
7	WC	RC	RC	RI	RC
33A	RN	RC	RC	WC	RN
36	RC	RC	WC	RN	RC
38	RC	RC	RC	RN	RC
45	WC	RC	WC	RN	WC
Knowledge of content and students items					
17B	RC	RN	RC	RN	RN
19	WC	RC	WC	WC	WC
25	WC	RC	WC	WC	WC
29	RC	RC	RC	RC	RC
30	RI	WI	WC	RN	WC

Note: R = Right; W = Wrong; C = Consistent; I = Inconsistent; N = Reason not explicitly stated and consistency could not be determined.

With regard to the kind of knowledge that teachers drew on to answer the questions (see Table 2), in only a handful of cases did teachers use guessing (2%) or test-taking strategies (1%). In just under a third of cases (32%) teachers used examples, counterexamples or pictures to justify (or attempt to justify) their responses. In roughly equal numbers of cases teachers drew on memorized rules or algorithms (13%), other mathematical reasoning (13%) and knowledge of students and content (12%). In 9% of cases mathematical justification was used and definitions were employed in 5% of cases. Like the consistent/inconsistent findings, these findings in relation to a small number of teachers in the pilot study support the inference in the elemental assumption that teachers' chosen responses to items were generally consistent with their mathematical reasoning about the items.

Table 2

Analysis of consistent-inconsistent responses in post-pilot test questionnaire

Item	Maria	Mark	Michael	Malcolm	Morgan
Specialized content knowledge items					
1A	NP	EX	KS	NP	EX
1B	EX	EX	EX	EX	EX
1C	NP	MR	NP	MR	EX
1D	EX	EX	MR	EX	EX
4A	MJ	OM	KS	MJ	MJ
4B	OM	EX	OM	NP	MR
4C	MJ	OM	OM	MJ	MJ
7	EX	OM	OM	EX	MJ
33A	NP	EX	MR	MR	NP
36	EX	EX	GU	KS	EX
38	EX	EX	EX	EX	EX
45	DE	DE	TT	DE	DE
Knowledge of content and students items					
17B	OM	NP	OM	NP	NP
19	KS	MJ	KS	KS	KS
25	MR	MR	GU	MR	MR
29	EX	EX	KS	OM	KS
30	EX	OM	EX	MR	NP

MJ, mathematical justification. Code reflects correct mathematical reasoning about an item (usually with reference to definitions, examples, counter-examples or unusual cases); MR, respondent refers to a memorised rule or algorithm; DE, captures inaccurate uses of definitions when responding to items. EX, examples, counterexamples or pictures. Use of numbers, cases, figures or shapes to assist in reasoning about an item. OM, other mathematical reasoning. Mathematical thinking not included in categories above; KS, knowledge of students and content. Respondent refers to knowledge of students to explain answer selection; GU, guessing acknowledged by respondent; TT, test taking skills used; NM, other non-mathematical thinking used; NP, not present. No reasoning was apparent from the transcript. (Codes taken from Hill et al (2007)).

3.2 The structural assumption

The inference of the structural assumption is that items will reflect the conceptual organization of the MKT theory, with regard to factors such as content knowledge (both CCK and SCK) and KCS. The organization of knowledge factors can be assessed using both exploratory factor analysis and confirmatory factor analysis.[8] Exploratory factor analysis identifies common factors among survey items without prior specification of factors. In a study such as this one confirmatory factor analysis has the advantage that hypotheses about factors derived from previous studies can be tested in a new country (van de Vijver & Leung, 1997, p. 99). I conducted both exploratory factor analysis and confirmatory factor analysis on the responses of 501 Irish teachers to survey items and expected to find that survey items were related to the hypothesized sub-domains of MKT. In other words, I anticipated that SCK items would load on one factor, CCK items on another and KCS items on another. The empirical findings of the exploratory factor analysis, however, provided little evidence to support the conceptualized categories (see Table 3).

Table 3

Promax rotated factor loadings with a three-factor solution based on data from Irish teachers

Item	Factor 1	Factor 2	Factor 3
C1(t)	**0.507**	0.041	-0.011
C2	0.281	0.252	0.061
C3	**0.634**	0.082	0.149
C4	**0.432**	0.117	0.329
C5	**0.418**	0.148	0.159
C6	**0.324**	0.131	0.158
C7	0.186	0.019	-0.015
C8	0.114	-0.165	0.074
C11	**0.400**	-0.121	0.232
C12	**0.531**	-0.188	0.033
C16	**0.428**	-0.114	0.331
C17	**0.407**	-0.226	-0.079
C18(t)	**0.618**	-0.017	-0.055
C19	**0.450**	0.044	-0.077
C20(t)	**0.335**	0.142	-0.122
C21	**0.358**	0.097	0.039
S9	**0.457**	0.139	-0.070
S10	-0.018	**0.309**	0.122
S13(t)	**0.520**	0.033	0.020
S14	0.233	-0.091	0.082
S15	**0.353**	-0.057	-0.082

S22	-0.031	0.164	0.064
S23	0.152	**0.940**	-0.118
S24	-0.051	0.052	**0.435**
S25	**0.348**	0.001	0.031
S26	**0.409**	-0.204	-0.022
S27(t)	**0.382**	0.289	0.056
S28	-0.007	0.137	**0.452**
S29	**0.334**	0.323	0.114
P30	0.193	-0.201	**0.544**
P31	0.019	0.047	**0.762**
P32	**0.500**	-0.047	-0.021
P33	0.049	-0.245	**0.531**
P34(t)	**0.578**	0.138	-0.001
P35(t)	**0.652**	0.201	-0.019
P36	**0.474**	-0.308	0.029

Bold print indicates the highest loading above 0.3 in a given row.

(t), testlet; C, content knowledge item; S, KCS item; P, algebra item.

Although initial analyses cast some doubts on the appropriateness of a three factor solution, I focused on such a solution because three factors were established in previous research (Hill et al., 2004). Contrary to expectations I identified one strong factor on which most content knowledge and algebra items loaded in the three factor exploratory factor analysis solution.[9] Seven KCS items loaded on the same factor. Two KCS items loaded on a second factor, and three algebra items and two KCS items loaded on a third factor. In summary, two-thirds of the items across three sub-domains loaded on one factor. Rather than three underlying factors explaining how Irish teachers responded to the items, this finding suggested that one strong factor, perhaps general mathematical knowledge, could explain teachers' performance on most items. These findings differed from factor analyses conducted on a parallel form in the United States[10] and reported by Hill, Schilling and Ball (2004). Correlations among the factors did not appear to be high (see Table 4).

Table 4

Correlations among factors in the three-factor, exploratory factor analysis model of the Irish teachers' data

	Factor 1	Factor 2	Factor 3
Factor 1			
Factor 2	0.091		
Factor 3	0.447	0.238	

I subsequently applied confirmatory factor analysis to the data. My goal in applying confirmatory factor analysis was to investigate if specifying the hypothesized factors in advance would provide greater clarity as to the factor loadings. In contrast to the exploratory factor analysis results, the confirmatory factor analysis indicated a clear algebra and a clear content knowledge factor (see Table 5). Nine KCS items loaded on a KCS factor. Confirmatory factor analysis produced better defined factors than exploratory factor analysis. One reason for the strong loadings in confirmatory factor analysis is that the factors [CK (made up of SCK and CCK), KCS and algebra] are strongly correlated among themselves. The correlations among the factors in the Irish data can be seen in Table 6. This suggests that rather than finding separate sub-domains of MKT, there appears to be one higher order factor, possibly MKT itself, which explains most of the variance among responses to items. Although the factors identified were different to those in the conceptual model of MKT, the factors - content knowledge, knowledge of content and students, and algebra - are broadly similar to those found by Hill et al (Hill et al., 2004) in the United States based on an adequate model fitting statistic.[11]

Table 5

Standardized confirmatory factor analysis for Irish teachers

	Irish teachers	
	Est.	SE
CK		
TC1	**0.489**	0.057
C2	0.356	0.058
C3	**0.717**	0.040
C4	**0.659**	0.063
C5	**0.571**	0.060
C6	**0.439**	0.056
C7	0.177	0.062
C8	0.126	0.062
C11	**0.513**	0.055
C12	**0.492**	0.054
C16	**0.588**	0.050
C17	**0.303**	0.063
TC18	**0.575**	0.037
C19	**0.430**	0.060
TC20	0.293	0.049
C21	**0.405**	0.057
KCS		
S9	**0.463**	0.055
S10	0.118	0.062
TS13	**0.549**	0.041
S14	0.259	0.062
S15	0.294	0.058
S22	0.043	0.063
S23	**0.317**	0.113
S24	0.250	0.066
S25	**0.376**	0.060
S26	**0.356**	0.064

TS27	**0.490**	0.045
S28	**0.315**	0.068
S29	**0.482**	0.068
ALGEBRA		
P30	**0.497**	0.073
P31	**0.550**	0.114
P32	**0.493**	0.063
P33	**0.341**	0.072
TP34	**0.664**	0.037
TP35	**0.729**	0.039
P36	**0.435**	0.065

Bold print indicates items which have a loading of 0.3 or higher

T, testlet; C, content knowledge item; S, KCS item; P, algebra item

Table 6

Correlations among confirmatory factor analysis factors in the Irish teachers' data

	CK	KCS
CK		
KCS	0.960	
Algebra	0.902	0.859

The other data that were available to assess the inference of the structural assumption were data from the pilot study interviews with five teachers (see Table 2). Among the KCS items just over a quarter (28%) were answered by teachers using their knowledge of students whereas knowledge of students was used by teachers just 5% of the time when responding to the SCK items. Although these results are in the direction that would be expected, they do not provide compelling evidence that KCS items draw primarily on knowledge of students.

The results of the factor analysis found among respondents to the items in Ireland are similar to the factors found among US respondents, suggesting that in both settings all items load on one strong (possibly) MKT factor. Despite the similarity of the factor analyses across countries, in relation to MKT in general, it suggests that the existence or perhaps the measurement of sub-domains may need to be reconsidered (also suggested by

Schilling et al., 2007, in relation to KCS, and CCK and SCK). The difficulty may be explained by items that poorly capture the hypothesized domains. Alternatively, if sub-domains of MKT exist, their specification may need to be reconsidered.

3.3 The ecological assumption

The inference of the ecological assumption refers to the extent to which teachers' scores on the measures were related to the mathematical quality of their classroom instruction. This was considered by studying the relationship between ten teachers' scores on the MKT items and the mathematical quality of the teachers' instruction, as coded using the MQI instrument. Because the ten teachers were selected as a convenience sample, there was a risk that the ten teachers would not be representative of the general teaching population. That concern was well founded. When the MKT scores of the ten videotaped teachers were considered alongside the 501 teachers who took only the multiple-choice measures (Delaney, 2008), in terms of their MKT the ten teachers in the videotaped sample ranged from the 36th to the 97th percentile of Irish teachers (see Table 7). In other words, all ten teachers are in the top two-thirds of Irish teachers, based on MKT scores. Furthermore, six of the ten are in the top quartile of Irish teachers. A wider spread of teachers along the MKT scale would have been desirable but recruiting such a spread of teachers would have posed practical and ethical problems. The relatively narrow range of teachers placed more demands on the MKT measures because they needed to be more sensitive to identify differences among teachers who are relatively close on the MKT scale.

The scale was developed using item response theory (IRT). A 2-parameter IRT model was made, using Bilog-MG version 3 IRT software (Zimowski, Muraki, Mislevy, & Bock, 2003), to estimate the likelihood of a teacher correctly responding to a multiple-choice question based on the teacher's underlying MKT. This model takes into account item difficulty and the fact that some items are better than others at predicting proficiency (Bock, Thissen, & Zimowski, 1997). Reporting raw scores on the measures would be problematic because items vary in difficulty, there is no expected performance level by which to judge teachers' scores, and some items are better at predicting teachers' overall MKT proficiency than others. Instead, the teacher scores are scaled to have a mean of 0 and a standard deviation of 1.0. A person with a score of 0 has a 50% chance of responding correctly to an item of average difficulty. Although the values can range from negative infinity to positive infinity, values typically lie between -3 and +3.

Table 7

The MKT score (range from -3 to +3) and MKT percentile of teachers in the video study, and their MQI global scores.

Teacher	MKT Score	MKT Percentile	MQI Global Score
Olive	1.879	97	3.53
Nigel	1.309	91	3.52
Brendan	1.279	90	4.51
Eileen	0.777	83	2.00
Clíona	0.677	82	4.76
Sheila	0.526	78	2.75
Veronica	0.357	57	1.72
Hilda	-0.141	46	2.72
Caroline	-0.357	42	2.77
Linda	-0.431	36	3.26

Percentiles calculated based on the MKT scores of all 501 teachers who participated in the MKT study

As mentioned above, coders[7] gave teachers a global score, estimating the teacher's mathematical knowledge as low, medium or high based on the MQI observed. In several cases coders chose intermediate levels of these bands (i.e. low-medium or medium-high) so in the analysis, a value was assigned to each lesson rating, from 1 (low) to 5 (high) with 2 and 4 representing intermediate levels.

Figure 3 presents a scatter plot of teachers' MKT scores and average global MQI scores over the four videotaped lessons. The regression line shows that in the case of five teachers - Caroline, Sheila, Hilda, Nigel and Olive - the MKT score was a good predictor of their MQI score. However, the MKT scores were not so good at predicting the scores of the other teachers. Clíona, Brendan and, to a lesser extent, Linda demonstrated a higher quality of MQI than would be expected from their MKT scores whereas Veronica and Eileen's MQI scores were substantially lower than their MKT results predicted. An examination of the correlation between MKT and MQI scores yielded no significant correlation between them. The absence of a significant correlation between teachers' MKT scores and their global MQI scores contrasts with a strong correlation found in similar analyses of US data (Hill, Blunk, et al., 2008). These results provided little support for the ecological inference. I now briefly consider possible reasons as to why the Irish results differed to those found in the United States sample.

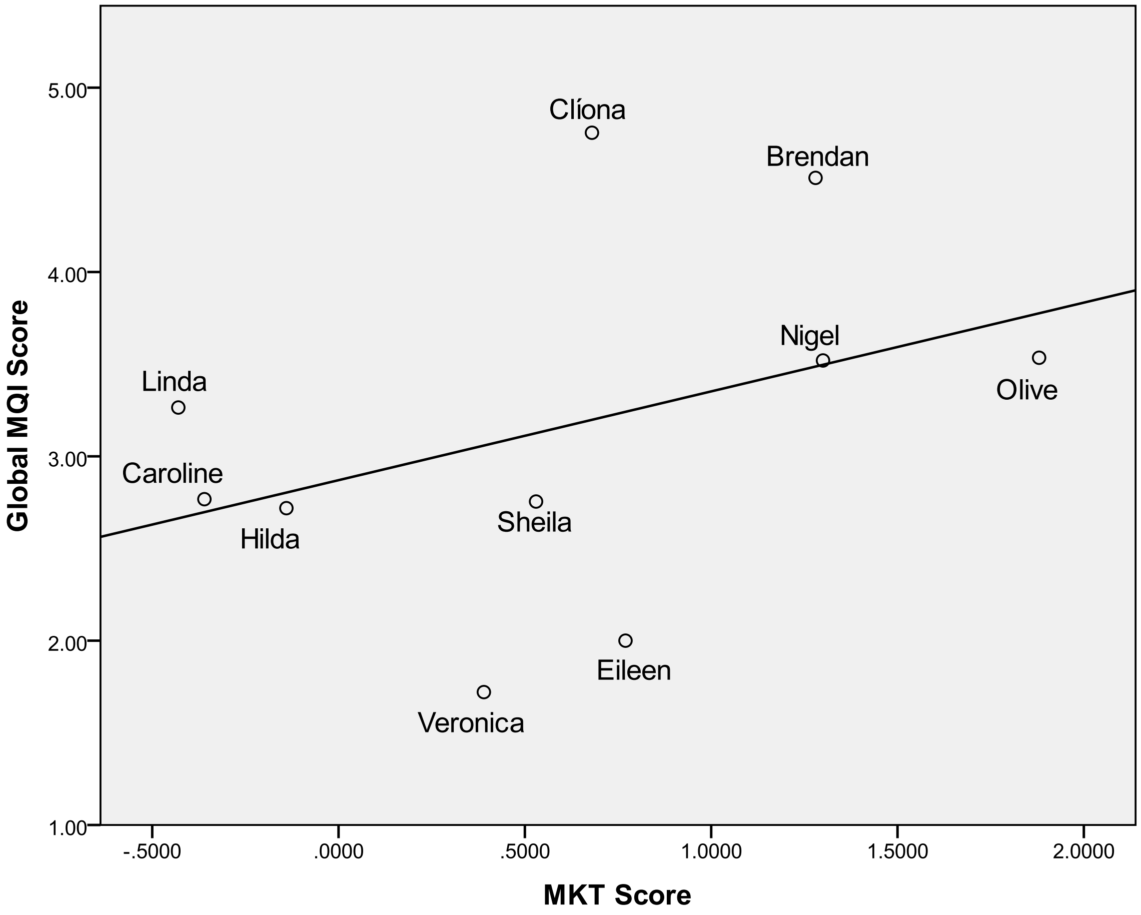


Figure 3. A regression line fitted to a scatterplot of teachers' scores on MKT and MQI.

The purpose of investigating the ecological inference was to determine the extent to which teachers' scores on the adapted MKT measures related to the mathematical quality of their classroom instruction. Finding a low correlation between the scores indicates that teachers' MKT is not strongly related to the teachers' MQI among this small, convenience sample of teachers in Ireland. The finding suggests that either the MKT items are not tapping into the mathematical knowledge that teachers use in practice or else that the MQI is not sensitively measuring the quality of mathematical instruction that was observed in these lessons. Before considering the implications of such a finding, I first look at possible reasons as to why the expected relationship was not observed in this sample.

First, the video study teachers were unevenly distributed on the MKT scale. Six teachers were in the top quartile of the population and no teacher was in the lower tercile of teachers. When teachers are located so close together on the scale, and when the items are poorly discriminating among them, the sample size is effectively reduced. Therefore, MKT scores and global lesson scores may be inconsistent due to measurement error. Because most teachers in the video sample scored highly on MKT, the lower performing teachers contribute most of the variance to the sample. But two of the lower performing teachers (Linda and Veronica) are outliers, in that one exhibited higher MQI than her MKT score predicted and one exhibited lower MQI than expected. Repeating this analysis with a set of randomly selected teachers would be desirable in order to investigate further the correlation between MKT and the MQI.

Second, the MKT test items were drawn from the strands of number, algebra and geometry whereas teachers in the video study were permitted to teach topics of their choosing. If the chosen topics were not well represented among the MKT items, and if the MKT items do not generalize across topics, this could have affected the relationship between the knowledge tapped in the items and the knowledge tapped in the MQI evaluation of teaching.

Third, teachers in the video study taught various grade levels from the equivalent of kindergarten to sixth grade. The knowledge that is tapped by the instruments may be more relevant for teaching mathematics to some grade levels than others.

4. Discussion

4.1 Evaluating the interpretive argument

I now return to my research questions and to evaluating the interpretive argument. First, the inference of the elemental assumption. Based on the pilot study interviews with five respondents, the teachers' thinking was consistent with their written responses on the multiple-choice items and most items were responded to using knowledge of mathematics or knowledge of students and very few by guessing or test-taking strategies. However, the sample size of five teachers and 17 questions was quite small. The inference of the structural assumption showed that the factors found among the items are similar to those found in the United States but the factor organization differs from the conceptualized domains of MKT. The ecological assumption was that teachers' scale scores on MKT measures were related to the quality of the teachers' mathematics instruction. A higher score was expected to be related to higher quality mathematics instruction and a lower score was expected to be related to lower quality mathematics instruction. Although the relationship existed in five of the ten teachers, half the sample consisted of outliers. Looking more closely at the videotape data, in the case of two teachers - Veronica and Eileen - it may have been because much mathematics class time was spent on non-mathematical activities, explanations were vague, and students' ideas were unchallenged, resulting in lower MQI scores than anticipated. Two teachers - Clíona and Brendan - demonstrated a higher level of MQI than expected. In one case this may have been achieved through detailed lesson preparation, an interest in language generally and by encouraging and challenging students.

Based on the findings of this validation study, the multiple-choice questions appeared to elicit the kind of thinking about mathematics and about teaching that was anticipated. However, further research is needed on how MKT is conceptualized and how well the measures are tapping into this knowledge. One way to do this would be for researchers to look at the characteristics of items that are better predictors of MQI than others. For example, higher correlations were found between MQI and MKT when the KCS items were excluded, and between MQI and MKT when only the algebra items were included (Delaney, 2008). Further validation studies of the measures in other countries, will help to elucidate our understanding of the MKT measures, complementing research in the United States, where researchers are considering revising the measurement of KCS and refining the specification for SCK (Schilling et al., 2007).

4.2 Challenges of validating measures of teacher knowledge

The study highlights the problematic nature of conceptualizing and measuring key constructs of professional knowledge in mathematics teaching. More specifically, it illustrates challenges in validating the use of test results when measures are adapted and transferred to a new setting. It may help explain why such validation is often neglected by test developers and users.

First, the need for further theoretical conceptualization of the domain of KCS items and the possible need for developing open-ended measures of teacher knowledge have been identified elsewhere (Hill, Ball, & Schilling, 2008). Given that the conceptualization of MKT is ongoing, it is difficult to know which aspects of the structural assumption are country-specific and which are more fundamental to the theory.

Second, the process is costly in terms of time and expertise. Even with the assistance of the MQI instrument that had been developed in the United States, users of the instrument need training in how to use it and if two people are to code instruction, at least one other person needs to be taught how to code.

Third, many factors may affect the correlation between MKT and MQI. The ten teachers taught different topics to different primary school age groups and this may have affected how they were ranked by their global MQI scores. Furthermore, the topics taught were not necessarily those tested by the MKT measures. For example, two strands of the Irish primary school curriculum – measures and data – were not explicitly covered by the MKT items. The tool used to analyse the MQI is not topic-specific. This meant that scores on the MKT measures were of necessity generalized to each teacher's overall MKT, and that the MQI evident in the four lessons taught was generalized to each teacher's overall mathematical instruction. Any inconsistencies in these generalizations may have contributed to noise in the correlation between MKT and MQI.

Fourth, the resources were not available to recruit a random national sample of ten teachers for the video study. Because a video study, where consent may be sought to show video at conferences or in professional development sessions, exposes a teacher's knowledge more than a written test, ethical concerns arise about deliberately recruiting teachers for the study who were likely to have a low MKT score. Nevertheless, the final sample and analysis would have benefited if more teachers in the sample had had lower MKT scores.

Fifth, the small sample size of ten teachers makes it difficult to get correlations between MKT and the MQI that are statistically significant.

Finally, one difference exists between using the MKT measures in Ireland and using the MQI instrument. Earlier I noted that the MKT measures were adapted for use with Irish teachers. In contrast, the MQI instrument was developed in the United States but was applied in a non-US setting. MKT measures need to be adapted because teachers' performance could be affected by measures which use terms that distract the teachers from the mathematical content of the measures. Furthermore, in using US measures in Ireland one must consider if the measures adequately represent the domain of practice in Ireland. For example, some measures may tap into knowledge that is not needed in Ireland and other knowledge that is needed in Ireland may not be measured by the items (for more see Delaney et al., 2008). The MQI instrument is different because it is an observational tool and it is applied after the teaching and cannot itself affect the teaching performance. It may, however, be incomplete if aspects of MQI that matter in Ireland are not part of the current codes.

Given that constructs developed in one country cannot automatically be applied in another country (e.g. Straus, 1969), validating the use of measures for new settings is essential. Over time this will need to be extended to measuring non-cognitive components of professional competence, such as beliefs and personal attributes (Blömeke, Felbrich, Müller, Kaiser, & Lehmann, 2008). As more experiences of validating measures and adapted measures of MKT are documented, researchers will benefit from using measures developed in different settings, and from analysing data gained about their own measures when they are used in new settings. By building on the work of documented validation work, test developers and users can become more sophisticated in validating the use of measures of teacher knowledge to inform education policy and practice.

Acknowledgments: This article is based on the author’s doctoral dissertation submitted to the University of Michigan. The author acknowledges his doctoral advisor, Deborah Loewenberg Ball for her support, as well as other members of his doctoral committee: Magdalene Lampert, Heather C Hill, and Kathleen M. Sutcliffe. He also acknowledges the contribution of Gabriele Kaiser and three anonymous reviewers for comments which substantially improved the article.

References

An, S., Kulm, G., & Wu, Z. (2004). The pedagogical content knowledge of middle school, mathematics teachers in China and the US *Journal of Mathematics Teacher Education, 7*, 145-172.

Andrews, P. (2011). The cultural location of teachers' mathematical knowledge: Another hidden variable in mathematics education research? In T. Rowland & K. Ruthven (Eds.), *Mathematical knowledge in teaching* (pp. 99-118). London and New York: Springer Science+Business Media.

Ball, D. L., & Bass, H. (2003). *Toward a practice-based theory of mathematical knowledge for teaching.* Paper presented at the annual meeting of the Canadian Mathematics Education Study Group, Edmonton, AB.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389-407.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133-180.

Blömeke, S., Felbrich, A., Müller, C., Kaiser, G., & Lehmann, R. (2008). Effectiveness of teacher education: State of research, measurement issues and consequences for future studies. *ZDM: The international journal on mathematics education, 40*, 719-734.

Blunk, M. L., & Hill, H. C. (2007). *The mathematical quality of instruction (MQI) video coding tool: Results from validation and scale building.* Paper presented at the American Educational Association Annual Conference, Chicago, IL.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*(3), 197-211.

Delaney, S. (2008). *Adapting and using US measures to study Irish teachers' mathematical knowledge for teaching.* Ph.D., Unpublished doctoral dissertation, University of Michigan, Ann Arbor, MI.

Delaney, S., Ball, D. L., Hill, H. C., Schilling, S. G., & Zopf, D. (2008). "Mathematical knowledge for teaching": Adapting US measures for use in Ireland. *Journal of Mathematics Teacher Education, 11*(3), 171-197.

Gorsuch, R. L. (1983). *Factor analysis* (Second ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methdos for test adaptations. *European Journal of Psychological Assessment, 11*(3), 147-157.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372-400.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430-511.

Hill, H. C., Dean, C., & Goffney, I. M. (2007). Assessing elemental and structural validity: Data from teachers, non-teachers, and mathematicians. *Measurement: Interdisciplinary Research and Perspectives, 5*(2), 81-92.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' knowledge for teaching. *The Elementary School Journal, 105, No. 1*, 11-30.

Hitchcock, J. H., Nastasi, B. K., Dai, D. Y., Newman, J., Jayasena, A., Bernstein-Moore, R., . . . Varjas, K. (2005). Illustrating a mixed-method approach for validating culturally specific constructs. *Journal of School Psychology, 43*, 259-278.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement: Fourth edition* (Fourth ed., pp. 17 - 64). Westport, CT: American Council on Education and Praeger Publishers.

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37*(2), 76-82.

Krauss, S., Baumert, J., & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: validation of the COACTIV constructs. *ZDM: The International Journal on Mathematics Education, 40*, 873-892.

Learning Mathematics for Teaching. (2006). A coding rubric for measuring the quality of mathematics in instruction (Technical Report LMT 1.06). Ann Arbor, MI: University of Michigan, School of Education. See http://sitemaker.umich.edu/lmt/files/lmt-mqi_description_of_codes.pdf. .

Learning Mathematics for Teaching. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education, 14*(1), 25-47.

Li, Y., & Even, R. (2011). Approaches and practices in developing teachers' expertise in mathematics instruction: An introduction. *ZDM: The international journal on mathematics education, 43*, 759-762.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.

Ma, L. (1999). *Knowing and teaching elementary mathematics*. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc.

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

Morris, A. K., Hiebert, J., & Spitzer, S. M. (2009). Mathematical knowledge for teaching in planning and evaluating instruction: What can preservice teachers learn? *Journal for Research in Mathematics Education, 40*(5), 491-529.

Muthén, L. K., & Muthén, B. O. (1998-2007). MPlus (Version 5). Los Angeles: Muthén & Muthén.

OECD. (2009). *PISA 2006: Technical report*. Paris: Author.

Organisation for Economic Co-operation and Development (OECD). (2004). *Reviews of national policies for education: Chile*. Paris: Author.

Organisation for Economic Co-operation and Development (OECD). (2008). *Reviews of national policies for education: South Africa*. Paris: Author.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Pepin, B. (2011). How educational systems and cultures mediate teacher knowledge: 'Listening' in English, French and German classrooms. In T. Rowland & K. Ruthven (Eds.), *Mathematical knowledge in teaching* (pp. 119-137). London and New York: Springer Science+Business Media.

Rowland, T., Huckstep, P., & Thwaites, A. (2005). Elementary teachers' mathematics subject knowledge: The knowledge quartet and the case of Naomi. *Journal of Mathematics Teacher Education, 8*(3), 255-281.

Schilling, S. G. (2002). ORDFAC software. Ann Arbor, MI: Author.

Schilling, S. G., Blunk, M. L., & Hill, H. C. (2007). Test validation and the MKT measures: Generalizations and conclusions. *Measurement: Interdisciplinary Research and Perspectives, 5*(2), 118-128.

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement: Interdisciplinary Research and Perspectives, 5*(2), 70-80.

Schmidt, W. H., Houang, R., Cogan, L., Blömeke, S., Tatto, M. T., Hsieh, F. J., . . . Paine, L. (2008). Opportunity to learn in the preparation of mathematics teachers: its structure and how it varies across six countries. *ZDM: The international journal on mathematics education, 40*, 735-747.

Schmidt, W. H., Tatto, M. T., Bankov, K., Blömeke, S., Cedillo, T., Cogan, L., . . . Schwille, J. (2007). The preparation gap: Teacher education for middle school mathematics in six countries (MT21 Report). East Lansing, MI: Center for Research in Mathematics and Science Education, Michigan State University.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57, No. 1*, 1-22.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: The Free Press.

Straus, M. A. (1969). Phenomenal identity and conceptual equivalence of measurement in cross-national comparative research. *Journal of Marriage and the Family, 31*(2), 233-239.

Stylianides, A. J., & Delaney, S. (2011). The cultural dimension of teachers' mathematical knowledge. In T. Rowland & K. Ruthven (Eds.), *Mathematical knowledge in teaching* (Vol. 50, pp. 179-191). London, New York: Springer Science+Business Media.

Tatto, M. T., Schwille, J., Senk, S. L., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher education and development study in mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics conceptual framework*. East Lansing, MI: Teacher Education and Development International Study Center, College of Education, Michigan State University.

Turner, F., & Rowland, T. (2011). The knowledge quartet as an organising framework for developing and deepening teachers' mathematical knowledge. In T. Rowland & K. Ruthven (Eds.), *Mathematical knowledge in teaching* (Vol. 50, pp. 195-212). London and New York: Springer Science+Business.

van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications, Inc.

Yang, X., & Leung, F. (2011). Mathematics teaching expertise development approaches and practices: Similarities and differences between Western and Eastern countries. *ZDM: The international journal on mathematics education, 43,* 1007-1015.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). Bilog-MG 3.0; Item analysis and test scoring with binary logistic models for multiple groups. Mooresville, IN: Scientific Software International.

Notes

[1] Middle school items have subsequently been developed.

[2] Four lessons were selected because in the US study four lessons per teacher was deemed to be the number of lessons needed per teacher to be safe in making inferences about the mathematical quality of teaching (Blunk & Hill, 2007).

[3] The teachers were actually asked about their answers to 16 questions but four of these questions were designed specifically for use in Ireland and because they were not adapted from US measures, they are excluded from this analysis.

[4] In total there are five sections and around 83 codes. Section 1 relates to instructional formats and content and section 4 relates to the textbook and teachers' guide. Codes from these sections will not be used in my analysis. In addition, the instrument has undergone modifications since it was used in this study. Details of the changes can be found at the website: http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&pageid=icb.page385579&pageContentId=icb.pagecontent818141&state=maximize (accessed on 29 January 2011).

[5] Coding for an entire lesson was rejected because the values of a particular code could vary across a lesson and it was difficult to recollect the entire lesson when coding; coding in ten minute segments was considered but rejected for similar reasons; five-minute segments were chosen but lessons were broken in the most natural point adjacent to the specific time.

[6] Research team members involved in the coding included teachers, teacher educators, and others. All have good knowledge of both mathematics and teaching.

[7] This process was followed for 70% of the Irish lessons and the remaining lessons were coded by the author alone.

[8] I acknowledge the assistance of Lingling Zhang and Laura Klem from CSCAR at the University of Michigan in conducting the factor analyses. Any errors are my responsibility.

[9] By convention, items are considered to load on a factor when the value is 0.4 or higher and 0.3 or higher when n > 175 (Gorsuch, 1983). In this case I used the criterion of > 0.3 to identify factors. In the Hill, Schilling and Ball (2004) study the criterion used was the highest loading on a factor.

[10] I used MPlus software, version 5 (Muthén & Muthén, 1998-2007), promax rotation and ULS (unweighted least squares) estimation. Hill, Schilling and Ball used ORDFAC software (Schilling, 2002) and promax rotation. No estimation method is specified.

11 The statistic used to establish model fit was RMSEA (Root mean square error of approximation), which describes the discrepancy between the data fit and a perfect fit. A measure of <0.5 is considered a good fit. The statistic for the model based on the Irish data was 0.027.