ZDM Mathematics Education (2012) 44:387–399
DOI 10.1007/s11858-012-0403-4

ORIGINAL ARTICLE

Analysis of psychometric properties as part of an iterative adaptation process of MKT items for use in other countries

Janne Fauskanger · Arne Jakobsen · Reidar Mosvold · Raymond Bjuland

Accepted: 28 March 2012 / Published online: 15 April 2012

Abstract Researchers at the University of Michigan have developed sets of items that can be used to analyze teachers' mathematical knowledge for teaching (MKT). In this paper, we consider what is required in the adaptation of a set of these items for use in a Norwegian context. We discuss how analysis of item difficulty and point–biserial correlation can be applied in combination with qualitative approaches to ensure a high-quality process of piloting adapted MKT items. Findings indicate that researchers who attempt to adapt MKT items for use in cultural contexts other than those for which they were designed need to use different methods to analyze all aspects of the adaptation process. The results from the different analyses conducted might then be used to inform other parts of the process, and this will mean that the process of adapting and piloting items becomes cyclic and iterative.

Keywords Assessment · Teacher knowledge · Mathematical knowledge for teaching (MKT) · Psychometric analysis · Cross-cultural adaptation · Item difficulty · Item development

J. Fauskanger (✉) · A. Jakobsen · R. Bjuland
Department of Education, University of Stavanger, 4036 Stavanger, Norway
e-mail: janne.fauskanger@uis.no

A. Jakobsen
e-mail: arne.jakobsen@uis.no

R. Bjuland
e-mail: raymond.bjuland@uis.no

R. Mosvold
Department of Early Childhood Education, University of Stavanger, 4036 Stavanger, Norway
e-mail: reidar.mosvold@uis.no

1 Introduction

Researchers have suggested that there is a connection between teachers' knowledge and the quality of their teaching (e.g., Darling-Hammond 2000; Hiebert and Grouws 2007; Hill, Blunk et al. 2008; Tchoshanov 2011). Several attempts have been made to describe the various components of this knowledge, and teacher knowledge is conceptualized using different frameworks (e.g., Askew 2008; Ball et al. 2001) and measured in various ways (e.g., Empson and Junk 2004; Hill et al. 2007). Shulman's (1986) distinction between aspects such as subject matter knowledge (SMK) and pedagogical content knowledge (PCK) has become famous in diverse areas of educational research (e.g., Graeber and Tirosh 2008). Within the field of mathematics education, Ball and her colleagues at the University of Michigan have contributed to a further development of Shulman's ideas. They have presented a framework for what they refer to as mathematical knowledge for teaching (MKT), and developed multiple-choice items to measure teachers' MKT (Ball et al. 2008; Hill et al. 2007). Results from their studies indicate that there is a connection between teachers' MKT and students' achievement in mathematics (Hill et al. 2005).

On the basis of the work of Ball and her colleagues, along with the apparent success of the efforts to measure vital aspects of teachers' MKT, we decided to implement the measures in a Norwegian context. Our starting point was to investigate whether and how the MKT measures could be used to study Norwegian teachers' MKT. Unlike the measures in, for example, PISA and TIMSS, the MKT measures were created on the basis of extensive studies of mathematics teaching in the USA, and were not intended for use outside of the USA. Since the knowledge required for teaching may be more culturally based than pertaining

simply to mathematical knowledge (e.g., Andrews 2011; Stylianides and Delaney 2011), attempts to adapt and use the MKT measures in a different cultural context should include careful analyses of the challenges involved.

Previous efforts have been made to investigate issues regarding translation, adaptation, and use of the MKT measures in countries such as Ireland (Delaney et al. 2008), Indonesia (Ng 2011), Ghana (Cole 2011), Norway (Mosvold et al. 2009; Fauskanger and Mosvold 2010) and South Korea (Kwon 2009). These researchers all approached the common problem of adapting an instrument devised in one cultural context to another that is different in various ways. This problem involves careful consideration of the content of the measures as they apply in different cultures, how the items represent the latent trait (construct) being measured (MKT), and how this construct is comparable across cultures. When adapting MKT items for use in other countries, these researchers have focused on documenting translation issues, interviewing teachers, and investigating psychometric properties (in particular item difficulty and point–biserial correlation) of the items. The process of adaptation has been limited to one cycle comprising translation, quality check, and finally a discussion of results from analyses of certain psychometric properties of the adapted measures. We focus on how analysis of psychometric properties can be used to inform the continued translation and adaptation of MKT items, with the following research question as focus:

> How can analysis of item difficulty and point–biserial correlation be used in combination with qualitative approaches to ensure an iterative and high-quality process of adapting MKT items for use in other countries?

By approaching this question, we aim to contribute to the understanding of the types of problems to be considered in translation, adaptation, and test score interpretation, and in particular how the different methodological approaches can become better connected. In this paper, we focus on the phase where items are adapted and piloted for use in a different cultural context, and our work is related in particular to the class represented by the Learning Mathematics for Teaching (LMT) measures. Our hope is that other research groups engaged in similar work can apply these principles to their own unique context and efforts to translate, adapt, and interpret the results of assessments of teachers' MKT.

2 Background

The study of mathematics teachers' knowledge has been an active field of research for decades (e.g., Sullivan and Wood 2008). Shulman's (1986) seminal work on the knowledge unique to teaching is frequently referred to (e.g., Graeber and Tirosh 2008), and his notions of SMK and PCK have been modified, subdivided, and refined. One of Shulman's (1986) recommendations for future researchers was to develop the knowledge base of teachers as well as tests which "those who have been professionally prepared as teachers are likely to pass…because they tap the unique knowledge base for teaching" (p. 13). One of the most widely recognized attempts to build on Shulman's recommendations is found in the LMT project (Ball et al. 2008), which includes a framework for teachers' knowledge base as well as measures for this knowledge. The LMT items[1] were intended to measure the MKT of practicing teachers, as opposed to, for example, the TEDS-M study where the focus was on pre-service teachers (Tatto et al. 2008).

2.1 The development of MKT items

MKT has been defined as "the mathematical knowledge used to carry out the work of teaching mathematics" (Hill et al. 2005, p. 373). As far as the development of teachers' knowledge base is concerned, the MKT framework splits Shulman's category of SMK into three sub-categories (e.g., Ball et al. 2008). Common content knowledge (CCK) refers to knowledge that is used in the work of teaching, in ways that correspond to how it is used in settings other than teaching. Specialized content knowledge (SCK), on the other hand, is defined as the mathematical knowledge "that allows teachers to engage in particular teaching tasks" (Hill, Ball, and Schilling 2008, p. 378). The third and final sub-category of SMK, knowledge at the mathematical horizon, is described as "an awareness of how mathematical topics are related over the span of mathematics included in the curriculum" (Ball et al. 2008, p. 403). Knowledge related to Shulman's PCK is divided into knowledge of content and students (KCS), knowledge of content and teaching (KCT), and knowledge of content and curriculum (see Fig. 1).

The MKT framework was developed through studies of different aspects of US teaching and relevant literature (e.g., Ball et al. 2008; Hill 2010), and the following initial question was raised: "What mathematical knowledge is needed to help students learn mathematics?" (Hill et al. 2004, p. 15). In order to consider how teachers' knowledge might be responsibly assessed, Hill and her colleagues (2007) aimed to move the debate concerning assessment of teachers "from one of argument and opinion to one of professional responsibility and evidence" (p. 112). In order

[1] From now on, for the sake of simplicity, we will refer to items from the LMT project as MKT items.

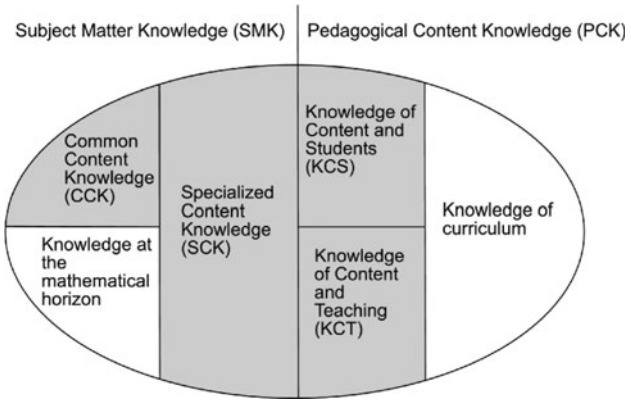

Fig. 1 Mathematical knowledge for teaching (Hill, Ball, and Schilling 2008, p. 377; *shading added*)

to make advances in developing tools to study teachers' knowledge, as well as to understand the MKT, a set of agreed-upon, reliable, and valid methods for assessing teachers' MKT is required (Hill et al. 2007). Assessment of teachers' knowledge must, according to Hill, Sleep, and colleagues, be strongly connected with the work of teaching, and they emphasize a further development of the MKT measures as one means of attaining such a goal. Studying the MKT items in cultural settings other than the USA is a natural follow-up. The domains of MKT have been tested and supported by psychometric analyses in the USA (e.g., Schilling 2007), but since the measures were developed on the basis of US teaching, they may not translate easily for use in other countries.

In 2001, Hill and colleagues started to develop multiple-choice items intended to represent elementary teachers' MKT (Hill et al. 2004). These items were developed for the following content areas: (a) number concepts and operations (NCOP); (b) geometry; and (c) patterns, functions and algebra (PFA).[2] The items were intended to represent the knowledge that is vital for teaching elementary mathematics. According to Hill and colleagues (2004), the initial item writing served a number of purposes. First, items were written to develop measures of teachers' MKT and to learn more about how MKT contributes to student achievement. Second, item writing was used to explore the nature and composition of MKT, and the very nature of MKT thus became strongly embedded in the items. Third, pilot testing of these items allowed the researchers to learn about the organization and characteristics of MKT. At present, items have been developed to measure teachers' knowledge in four of the MKT domains: CCK, SCK, KCS, and KCT (Hill 2010), shaded in Fig. 1.

[2] See Ball and Hill (2008) for a sample of released items.

The researchers have so far been more successful in developing items to measure the SMK domains (i.e. CCK and SCK) than the PCK domains (Hill 2010).

By nature, the MKT items are closely connected with teaching practice, and researchers such as Stigler and Hiebert (1999) have argued that there are cultural differences in teaching practice. The issue of cultural differences in teaching is complex, however, and it has been argued that there are both differences and similarities between countries (Anderson-Levitt 2002). The developers of the MKT items have focused particularly on challenges that are specifically related to the work of teaching, and there appears to be an underlying assumption that these challenges (referred to as "tasks of teaching") are similar across countries (Ball et al. 2008).

A number of mathematical tasks of teaching have been identified in the USA. Two examples are presenting mathematical ideas and choosing and developing usable definitions. On close examination, though, some of these tasks may be foreign to teachers in other countries. If the tasks of teaching differ, then the MKT is also likely to differ. Kawanaka et al. (1998) found that even though teachers in different countries are involved in the same activities, "there were enormous differences in how those activities were done" (p. 93). These differences may influence the mathematical knowledge required. Thus, it is important to question whether or not the demands for mathematics teaching in other countries are similar to the knowledge conceptualized in MKT.

The MKT items were written to be specific to issues of context and they are thus subject to cultural variability in teaching and schooling. This is the primary reason why translating these items is more challenging than translating a mathematics test for students, but this is also a feature that makes it more interesting. Attempts to adapt and use the MKT measures in a different cultural context should include careful analysis of the challenges involved on different levels. As an example, prior research on US teachers' subject-matter knowledge found that many teachers hold procedural understandings of algorithms, in contrast to teachers in China (Ma 2010). When adapting an item focusing on, for example, algorithms, which might differ across countries, this is an important issue to take into consideration.

2.2 Translating and adapting MKT items

In a study in which MKT items were adapted for use in Ireland, Delaney and colleagues (2008) discussed cultural differences extensively. They focused on the process of item adaptation in particular, and they started by documenting all changes that were made to the items. These changes were divided into the following categories: (1)

changes related to the general cultural context; (2) changes related to the school cultural context; (3) changes related to mathematical substance; and (4) other changes. As part of their study, they also evaluated the adapted items through interviews with Irish teachers, and they analyzed and compared item difficulty and point–biserial correlation in Ireland and the USA.

Building on the efforts of Delaney and colleagues (2008), several researchers have adapted MKT items for use in other countries. Documentation of translation and adaptation has been done in Ghana (Cole 2011), Indonesia (Ng 2011), Norway (Mosvold et al. 2009), and South Korea (Kwon 2009). All of these studies made use of some kind of focus-group to provide quality assurance of the translation. Only Kwon (2009) used back-translation to ensure the quality.

In Ghana (Cole 2011) and Norway (Fauskanger and Mosvold 2010), follow-up interviews were conducted with teachers who had been measured. Cole (2011) also conducted studies of Ghanaian teachers' mathematical quality of instruction. In both the Ghanaian and Norwegian cases, challenges regarding the item format were analyzed. Analysis of item difficulty as well as point–biserial correlation were carried out in all of the studies, and Kwon (2009) used these analyses of the item's psychometric properties to discuss translation issues. Her discussions did not, however, appear to result in continued adaptation of the items.

The present article is an attempt to continue these discussions of adapting MKT items for use in different contexts. Building on experiences from the previous studies in this area, we examine how analysis of item difficulty and point–biserial correlation can be used to inform the qualitative approaches in a continuing adaptation process of MKT items.

3 Methods

In the first phase of our project, a complete form[3] of items from the LMT project was translated and adapted for use among Norwegian teachers (Mosvold et al. 2009). This form contained the following three MKT scales:[4] NCOP (27 items); geometry (GEOM, 19 items); and PFA (15 items). Most forms contained only one content area, and we selected this one because it contained three areas that are emphasized in the Norwegian curriculum. In total, 30 item stems and 61 items were included.

[3] Elementary form A, MSP_A04.

[4] We use "scale" when referring to sets of items within a form.

3.1 Translating items

Since single translation of items has proven to be the least trustworthy method, we decided to use double translation (Adams 2005), which means that two independent translations are made from the source language with reconciliation by a third person. In our project, pairs of researchers with a specialty in mathematics education made two parallel and independent translations. These two translations were then compared and discussed, and a final translation was made. Throughout the translation process, all changes that were made to the items were documented using the four categories from Delaney and colleagues (2008). Due to some particular challenges that occurred in the Norwegian context, however, the list of categories was developed further and two new categories were included: (1) changes related to the translation from American English into Norwegian in this particular context; and (2) changes related to political directives (Mosvold et al. 2009). The first category replaced an original sub-category concerning changes related to spelling in the first category from Delaney and colleagues (2008). Translation into a different language goes beyond differences in spelling, and translating the MKT items into a different language is complex. The second category was added in order to cover some issues related to directions that Norwegian schools received from the Ministry of Education and Research (see Mosvold et al. 2009).

3.2 Focus-group interviews

Seven semi-structured focus-group interviews (FGIs) were organized with 15 participating teachers. Teachers from different schools and grade levels and with different levels of experience were selected. The participants in the first two groups were selected on the basis of their level of experience and special interest in mathematics education, and all taught at different schools. The first group consisted of two experienced teachers, while the second group consisted of three inexperienced teachers. The other five groups were randomly selected from schools that were connected to our university as practice schools for preservice teachers in collaboration with their respective headmasters. All the participants had a special interest in mathematics and mathematics teacher education.

The FGIs were conducted after the teachers had worked individually with the items in a testing situation. The interviews focused on five groups of questions. First, a group of questions related to background information of the teachers. Second, we asked general questions about the MKT measures, such as their views of the relevance of the items' context and more general comments about the measure as a whole. Third, we asked if the teachers had

comments related to the format of the items. Fourth, we asked them to comment on the mathematical topic, structure, and difficulty item by item. Finally, we asked them to supplement the other comments and reflections discussed in the interview. The FGIs were recorded and transcribed, and these transcriptions were analyzed through content analysis (e.g., Törner et al. 2010). The aim of content analysis is "to obtain descriptive information about a topic" (Fraenkel and Wallen 2006, p. 485). For the purpose of this paper, the transcriptions were analyzed according to how the teachers commented on each individual item.

3.3 The adapted MKT measures

After having translated and adapted the items, 142 teachers' MKT were measured. The participants were selected from a convenience sample (Bryman 2004) of 17 schools, all of which were connected to our university as practice schools. Among the participating teachers, 41 worked in grades 1–7 and 96 in grades 8–10. Since Norwegian teacher education was previously similar for teachers from grades 1–10,[5] the teachers had the same formal qualifications even though they taught different grade levels. Five teachers did not provide any information about their teaching.

In order to learn more about how the adapted measures functioned in the Norwegian context, we conducted psychometric analysis by applying a two-parameter logistic model (2PLM) on our data.[6] The 2PLM is one among a number of item response theory (IRT) models, and it was chosen because it was the model used in the LMT project for analyzing the results for the same form we used in this study (Hill 2007). IRT models are not sample-dependent, and they are fairly robust in estimation of item difficulty and discrimination when a convenience sample is used, as in this case (Alexander 1992). It is often recommended that one should have at least 200 respondents when using a 2PLM (Edwards 2009), and since we had a lower number of respondents, we had to be careful about how we should interpret the results. As in the LMT project (Hill 2007), we used BILOG-MG (Zimowski et al. 2003) for the estimation of item difficulties and point–biserial correlation. Our main focus when analyzing these psychometric properties was on identifying items with negative point–biserial correlation and comparing relative item difficulty estimated in Norway and the USA.

[5] In 2010 the Norwegian teacher education changed and was divided into grades 1–7 and 5–10.

[6] Missing data is not used in parameter estimation in the models, and we did not correct for the violation of local independence by testlet items.

It is assumed that a teacher's ability to answer an item correctly is a function of both person properties and item properties (Edwards 2009). This relationship is modeled according to the item characteristic function (or item characteristic curve). The teachers who are measured are assumed to have a numeric value that places them somewhere along the ability interval, which is typically from −3.0 to +3.0, with 0 as the mean ability level and a standard deviation of 1.0. The probability that a teacher can answer each item correctly will be low for those who have low ability and high for those with high ability. This relationship is portrayed for three test items in Fig. 2.

Each item has what is called a difficulty parameter, which is the point along the continuum where an individual has a 50 % chance of correctly answering it. A 2PLM allows for items with different item slopes, which describe how well an item discriminates among teachers. Due to the relatively low number of participants in our study, we focus our discussions on item difficulty and not on slopes. We calculated the Pearson's correlation coefficient to investigate the correlation between relative item difficulty found in Norway and the USA.

If adapted scales have not been calibrated against the original scales, item difficulties are not directly comparable without further analysis of raw data from both countries. Since US raw data were not available, it was not possible to do this. Instead, we take a more exploratory approach and focus on relative ordering of item difficulty in our analysis. If, for instance, adapted measures consist of items that are relatively easier for Norwegian teachers than for teachers in the USA, the measures would be less useful than intended. After having calculated the item difficulty for all the items in our adapted form, we ordered the items according to their relative item difficulty in both countries. Ordering is the ranking of items according to relative item

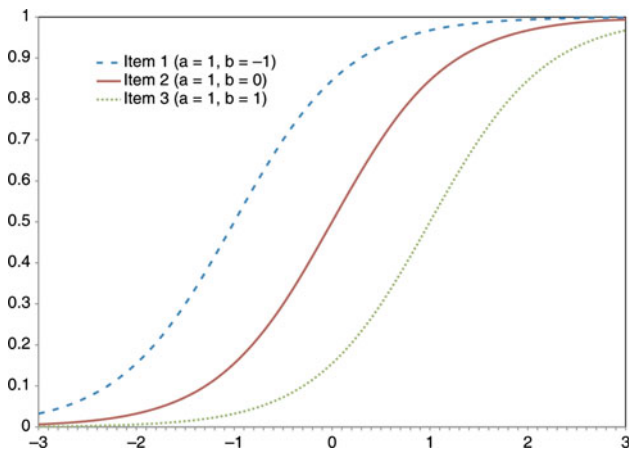

Fig. 2 Item characteristics functions/curves for three items with the same slope (*a*), but different item difficulty (*b*) (Edwards 2009, p. 510)

difficulty from easy to hard; a change in ranking of an item is related to a different item difficulty ordering in the other country.

Point–biserial correlation coefficients were also calculated in BILOG-MG, and they tell us how strongly individual items are correlated with the rest of the items. High point–biserial correlation for an item indicates that there is a strong relationship between that item and the underlying construct being measured (Delaney et al. 2008; Harvey and Hammer 1999). A negative point–biserial correlation indicates that respondents who answered other items correctly would probably give the wrong answer to this item, and such items should receive particular attention since they might have to be discarded from the measures (de Ayala 2009).

4 Results

The focus in this paper is on how analysis of relative item difficulty and point–biserial correlation can be used in combination with the qualitative documentation of the translation process and FGIs in order to ensure an iterative and high-quality adaptation process of MKT items. In this section we present some of the results from our psychometric analyses as well as from our FGIs and process of translation.

4.1 Psychometric analyses

Previous analyses have already indicated that the adapted Norwegian measures function well overall (Jakobsen et al. 2011). Reliability estimates are good for all three scales (Table 1). As can been seen from Table 1, the point of maximum information is below the mean for all three scales, indicating that all scales provide optimal measurement of teachers who are less knowledgeable than the average (from 0.75 to 0.875 standard deviation below the mean score).

When analyzing difficulty, we revealed that item difficulty found in Norway was distributed over the ability interval −3.432 to 2.534, whereas in the USA it was over the interval −3.734 to 3.454. The distribution of item difficulty for each scale is shown in Table 2, together with the average item difficulty on each scale.

Table 1 Reliability estimates and points of maximum information for all scales

	Number of items	IRT reliability	Max information
NCOP	27	0.838	−0.8750
GEOM	19	0.799	−0.7500
PFA	15	0.861	−0.7500

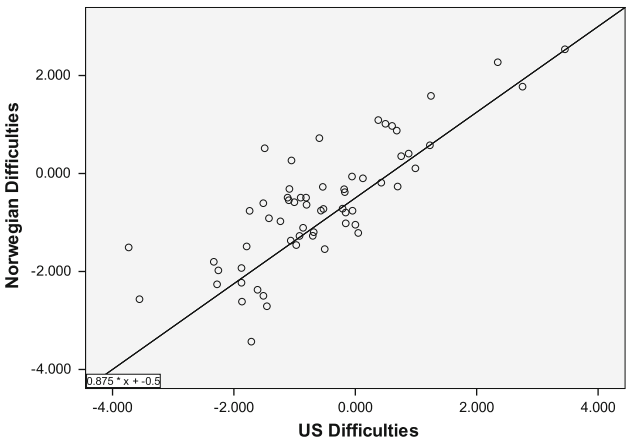

Fig. 3 A scatter plot of the relative difficulty for items found in Norway plotted against US difficulty

A scatter plot showing Norwegian item difficulty plotted against the difficulty found for the same items in the USA is shown in Fig. 3. We see from the scatter plot that most items are located close to the regression line, which indicates that the relative item difficulty is similar for the items in the two data sets. The estimate of item difficulty for each item also has an associated standard error, and in our adapted measures, the standard error of the item difficulty ranged from 0.102 to 0.937, with a mean standard error of 0.298 and a standard deviation of 0.182. We find a strong correlation between the difficulty estimated in Norway and what is reported in the USA (Pearson correlation is 0.812, $p < 0.0005$), and the average item difficulty is similar (−0.57 in the USA vs. −0.64 in Norway). From this we conclude that there is a strong relationship between relative item difficulty in Norway and the USA, allowing us to identify sizable (relative) differences.

Due to the way the item parameters were estimated, they are not directly comparable and we had to rely on the relative differences in the ordering of the items according to item difficulty in the two countries. We decided to further investigate items that had a greater than one unit change in difficulty ordering between the Norwegian and US samples[7] (Table 3). This is much higher than can be explained by error estimates, and is the situation for items 6, 11a, 11b, 14a, 15a, 17b, 21, 25a, and 28.

By analyzing the point–biserial correlations for the adapted measures, we found that only one of the items

[7] As noted by one of our reviewers, there are several ways one could look for items that appear to be behaving differently in the two samples. One suggested method, comparing the rank of items by their difficulty in the two different calibrations, was also considered. The ranks, while acknowledging that the metrics for the difficulties are not necessarily the same, can show extreme changes in areas where lots of the difficulty values cluster. Ultimately, we decided to focus on the difficulties of items, but acknowledge that there are other legitimate choices that may result in different conclusions.

Table 2 Number of items within different ability intervals and average difficulty for all three scales

Difficulty	<−2	[−2, −1⟩	[−1, 0⟩	[0, 1⟩	[1, 2⟩	>2	Average Norway	Average USA
NCOP	2	7	10	3	2	2	−0.389	−0.461
GEOM	5	6	4	3	1	0	−1.086	−0.528
PFA	1	2	9	3	1	0	−0.517	−0.808

Table 3 Items with greater than one unit change in difficulty

Item	US difficulty	Norwegian difficulty	Change in difficulty (absolute value)
14a	−3.734	−1.509	2.225
25a	−1.493	0.514	2.007
17b	−1.713	−3.432	1.719
28	−1.052	0.266	1.318
6	−0.593	0.722	1.315
11a	0.048	−1.213	1.261
15a	−1.457	−2.709	1.252
21	0.001	−1.047	1.048
11b	−0.504	−1.543	1.039

stands out. Item 17c, among the geometry items, had a negative point–biserial correlation of −0.053 in our adapted measures. This item belongs to testlet[8] number 17, in which the respondents were supposed to figure out which object descriptions are possible and which are not. The definition of a parallelogram based on the length of the diagonals was the focus in this item.

Based on the findings from our analysis of point–biserial correlation as well as comparison of ordering of items according to item difficulty, we can divide the items in our adapted measure into three groups:

1. Items that do not seem to function in Norway ($n = 1$, item 17c).
2. Items that function well, but have a relatively high difference in item difficulty ordering in Norway compared with the USA: 6, 11a, 11b, 14a, 15a, 17b, 21, 25a, and 28 ($n = 9$).
3. Items that seem to function well and that have item difficulty ordering close to what is reported in the USA ($n = 51$).

The items in the third group seem to function well, and we decided to select items from the first two groups for further analysis in this paper (see Table 4). Group 1 only contains one candidate (item 17c). From group 2, we discuss two items. Item 6 was ordered on the easy side in the USA (difficulty −0.593), while the difficulty in Norway was 0.722. This shift in difficulty was a reason why we decided to conduct a more careful analysis of the item. Items 25a and 28 had a similar shift in relative item difficulty, but we decided not to focus on these two items in this paper. They were both from the PFA scale and some teachers were concerned about being tested in areas they did not regard as relevant for their teaching (Fauskanger and Mosvold 2010; Jakobsen et al. in press). Item 6 was frequently mentioned in the FGIs we conducted as part of our adaptation of the set of MKT measures (Fauskanger and Mosvold 2010), and this also made it interesting to discuss. Second, we consider item 14a more closely. This item had the greatest difference in difficulty ordering of all items (difference in difficulty 2.225).

[8] A testlet has one item stem with several related items below.

4.2 Focus-group interviews

After having identified these problematic items, we went back to the transcriptions of the FGIs to investigate how the teachers had reacted to these items. Concerning testlet 17, we found that most teachers commented on this testlet. From their comments, we conclude that teachers find items requiring definitions of polygons in general, as well as hierarchical definitions, difficult. The teachers claim that they rely on the definitions presented in the textbooks and that they use these if definitions are needed.

Some of the nine items that stood out with a relatively high difference in item difficulty ordering also stood out in the FGIs. Item 6 represents an example of this. It seems as if this way of discussing the multiplication of fractions and representing it on the number line was unfamiliar for Norwegian teachers (Fauskanger and Mosvold 2010). In six of the seven interviews, this item was considered to be problematic. From the teachers' explanations in the FGIs, it seems as if the context was unfamiliar and difficult to understand, and the problems might be related in particular to the representation of mathematical ideas in the item. The excerpt from one of the interviews below illustrates the confusion:

131. Interviewer: Item 6?
132. Teacher 13B: You know what? I simply had to skip that one, because I
(…) I didn't have a clue! So I came back to it in the end (…)
133. Teacher 13A: I used the common denominator on that one…
134. Teacher 13B: I think it was a confusing question, I mean, that they should jump along the number line (…) (Transcriptions, school 13)

Table 4 Items to be discussed

Item number (scale)	Content	Context
6 (NCOP)	Multiplication of fractions represented by a frog's movement on a number line	The students have been given a set of directions to move the frog on a number line but do not agree at which point the frog will stop. The teachers are asked to mark which of the students' answers they should accept as correct
14a (NCOP, testlet)	Whole number divided by a fraction. In 14a the students' solution is not the one the teacher presented in the item stem expected, but it is a correct solution	The students have been given a task and four different student solutions are presented. The teachers are asked to evaluate each of the four solutions and to figure out which solutions are valid
17c (GEOM, testlet)	Definitions of polygons. In 17c the focus is on the definition of a parallelogram based on the length of the diagonals	The students are asked to make descriptions of polygons that do not exist. The teachers are asked to judge four such descriptions

With respect to the content of testlet 14, the teachers agreed that it was an interesting problem. In the FGIs, the teachers also said that they were not used to evaluating students' written answers in this way and found the items in this testlet difficult (132). The discussion from the interview shows that these teachers found it both difficult and confusing (132, 134), and they were confused by the context where the frog was jumping along the number line (134). The Norwegian teachers would have asked the students about their thinking in order to understand and assess their answers rather than figuring out what the students were thinking from their written work. In each of the items in testlet 14, the teachers have to evaluate students' suggested solutions, and being unfamiliar with this kind of evaluation may complicate the item. Several interviews also contained discussions of the word "valid" and whether a valid solution means that the students' solution needs to be correct or not. Some teachers interpreted our translation of valid as "on the right track" and not necessarily mathematically correct. In this item "valid" means "mathematically valid", and the teachers' perception may thus lead them to misinterpret this.

4.3 Translation of items

Based on our analysis of the items' point–biserial correlation, we found that item 17c did not work properly in the Norwegian study. When translating this entire testlet, the following kinds of changes were made (Table 5):

Only minor changes related to language differences were made in item 17c. For the item stem of testlet 17, however, we made changes in all the other categories.

Item 6 was related to the multiplication of fractions, and it appeared to be a rather unproblematic item to translate. We only made corrections of general cultural context by changing the teacher's name from Ms. Lee to Kari (which is a common Norwegian first name), and we also changed "group of…students" to the Norwegian equivalent of "pupils". After having observed how this item stood out in the analysis of item difficulty, however, we went over the translation again and discovered a phrase that could be misinterpreted in the adapted item. The phrase "set of directions" from the original item was translated into Norwegian in a way that is closer to "set of rules". Since "rules" has a specific meaning in mathematics, some teachers may have found this confusing and tried to recall what mathematical rules they were supposed to use in item 6. This again highlights the importance of good translation.

Based on our comparison of relative item difficulty above, item 14a also appeared problematic. Whereas item 14 was on the whole rather complicated to translate, part (a) of the item did not contain any particular challenges of translation. In addition to the more common change of names, the item stem included a context that is uncommon in the Norwegian context. The item was related to division of fractions, and the context of the problem referred to chocolates that were bought in the school candy sale. In Norwegian schools, a school candy sale is uncommon and

Table 5 Changes documented for testlet 17

Changes related to the general cultural context	teacher's name changed from Mr. Erikson to Håkon (a common Norwegian first name)
Changes related to the school cultural context	"students" changed to a Norwegian equivalent of "pupils"
	"class discussion" changed to a similar Norwegian term where the word "class" is not used
Changes related to the mathematical content	"polygon" changed to a Norwegian word that literally means "multi-edge"
	"right triangle" changed to a word meaning "right-angled triangle"
Changes due to language differences	"of equal lengths" changed to a Norwegian phrase meaning "has equally long [diagonals]"

it was therefore necessary to rewrite this somewhat. We decided to translate "school candy sale" with "butikken" (which is the Norwegian word that would be used to describe a generic food/grocery store). Some changes were also made in relation to the mathematical language, and the words "add" and "subtract" were replaced with the more informal Norwegian "legge sammen" (lit.: put together) and "trekke fra" (lit.: take away from). All of these changes were, however, made in parts other than the relevant part (a) of this testlet item.

5 Discussion

In previous publications we have discussed results from our process of translating and adapting items (Mosvold et al. 2009; Ng et al. 2012) and analysis of teachers' responses from the FGIs (Fauskanger and Mosvold 2010), and have also presented a more limited quantitative analysis of the results (Jakobsen et al. 2011). In the present paper we have used the analysis of relative ordering of item difficulty and point–biserial correlation to uncover potentially problematic items, and we have then gone back to discuss the translation, adaptation and teachers' responses to some of these particular items all over again.

5.1 Item 17c

The raw score for item 17c was relatively high, and it came as a surprise that the point–biserial correlation coefficient indicated that this item did not function. When going back to the teachers' responses in the FGIs, we found that in all the interviews teachers commented on item 17 in general. They seemed to find items requiring mathematical definitions difficult. Unfamiliarity related to definitions may therefore be an explanation of why this item did not function in Norway. On the other hand, the other three items (17a, b, and d) of this testlet function well, and all of them are based on definitions. The problem might therefore be caused by the way item 17c focused on defining the parallelogram by considering the length of the diagonals. This indicates that the task of choosing and developing usable definitions (Ball et al. 2008) may not be identical in Norway and the USA. We find it interesting to observe that Ng (2011) reported a negative point–biserial correlation of −0.045 for this exact same item in the Indonesian context. One of his explanations for this was that Indonesian textbooks and curriculum materials do not focus on this kind of definition, and teachers might thus lack knowledge of how different geometric objects are related. All the items that measured teachers' content knowledge of the hierarchical relationships of quadrilaterals were more difficult for Indonesian teachers than for those in the USA (Ng 2011). Looking back at the process of translation and FGIs does not indicate that the reason for this item's dysfunction was related to translation and adaptation, and there is thus a possibility that this might be true for the Norwegian context as well.

5.2 Item 6

Item 6 was from the NCOP scale. The item had a relatively high difference in item difficulty ordering, and the item displayed a shift in relative item difficulty when adapted and used in a Norwegian context. In the USA, the item was ordered slightly on the easy side (with a difficulty of −0.593), while the difficulty was 0.722 in Norway. The item presented a context where multiplication of fractions was represented by a frog's movement along the number line. This way of talking about multiplication of fractions and representing it on the number line was unfamiliar for the Norwegian teachers (cf. Fauskanger and Mosvold 2010). When going back to analyze the teachers' comments in the FGIs, we found that the item was considered problematic by the teachers in six of the seven FGIs. The Norwegian teachers found the context unfamiliar and difficult to understand, and the reason for the difference in difficulty ordering may be related to differences in how mathematical ideas are represented in Norway and the USA.

In the process of translating item 6, some changes were made in relation to the general cultural context as well as to the school cultural context, but we did not initially find the item particularly problematic to translate. After having observed how this item stood out in the psychometric analyses, however, we discovered a problem with the translation of "set of directions" to a phrase closer to "set of rules". This discovery demonstrates the importance of using the analysis of the psychometric properties of the items to uncover potentially problematic items. It also clearly demonstrates how the different methodological approaches can and should be analyzed in connection with each other, and indicates that the adaptation of items needs to be seen as an iterative process.

5.3 Item 14a

Another example of an item with high difference in item difficulty ordering is item 14a. This item had negative difficulty in both countries (−1.509 in Norway and −3.734 in the USA), and it was thus a fairly easy item in both countries. It was a challenging item to translate, however, and we argue that the reason for the relatively high difference in item difficulty may be related to translation. When the comparison of item difficulty ordering indicated that this item might have problems, we went back to

analyze the translation and adaptation of the item. In this new analysis, some other problematic aspects with the translation were revealed. The item stem contained some words (e.g., "valid") that were replaced by more informal words in the adapted item. We decided, for instance, not to use the Norwegian equivalent of "valid" (also written "valid") in our adapted item, but rather replaced it with the term "holdbar", which directly translates to "durable". In the FGIs, some teachers revealed that they found this item difficult, with one reason given being that the term "holdbar" can be interpreted to denote a weaker status than "valid". Again, we see how the analysis of psychometric properties prompted us to look at the other sources of data with new eyes.

In four of the FGIs teachers said that they were not used to evaluating students' written answers as presented in testlet 14. The teachers would ask the students about their thinking in order to understand and assess their solutions rather than figuring out what the students were thinking from their written work. In each of the items in testlet 14, the teachers have to evaluate students' suggested solutions, and lack of familiarity with this kind of evaluation may complicate the item. From the FGIs, we see that the amount of text given in the different MKT items is problematic for the teachers (Fauskanger and Mosvold 2010). They are not used to reading a lot of text in relation to mathematical items, especially text including mathematical concepts, which is the case in testlet 14 and the related items. This may be part of the reason why item 14a has a high difference in item difficulty ordering, but it does not explain why 14a differs from 14b, c, and d. Therefore, the reason may not be connected to the amount of text. It may, for instance, be a matter related to the length of the test rather than the wordiness of individual items.

The teachers also said that there are too many items focusing on fractions in our measures compared with how much emphasis the topic is given in Norwegian classrooms. Part of the explanation of why a high difference in item difficulty ordering exists in item 14a might therefore be found in cultural differences in the work and tasks of teaching. On the other hand, there are several items related to fractions, so this does not explain the difference in item 14a in particular.

5.4 Summing up

We identified nine items that function well but have a relatively high difference in item difficulty ordering in Norway compared with the USA. When studying the other seven items in this group (1a, 11b, 15a, 17b, 21, 25a, and 28), we find that challenges related to item 6 and 14a give a true picture of challenges related to the other items.

Other researchers have analyzed different aspects regarding the adaptation of MKT items for use in other countries, and most of them build on the efforts of Delaney and colleagues (2008). The importance of documenting and analyzing the process of translating the items has been emphasized (e.g., Mosvold et al. 2009), and efforts have even been made to discuss the connection between analysis of psychometric properties (in particular item difficulty and point–biserial correlation) and the translation process (e.g., Kwon 2009). These latter efforts have so far been confined to theoretical discussions of how the translation might have influenced the psychometric properties. The results from an analysis of psychometric properties have thus been discussed in relation to an already completed adaptation process. In this study, we have taken this one step further by showing how analysis of item difficulty and point–biserial correlation can be used to illuminate problematic items, which then leads to another cycle of revision in the process of adapting the MKT items. Even if the IRT-reliability in our study was good for all scales—unlike what was found to be the case in Indonesia (Ng 2011)—our results indicate that the process of piloting adapted items might need to continue for several cycles. Figure 4 illustrates a summary of our efforts and indicates that a high-quality adaptation process of the MKT items should be a cyclic process.

The figure also indicates that, when analyzing results from using the adapted measures, one may need to go back to the cyclic process of adaptation in order to make sure that the results are not biased due to an adaptation process of low quality (indicated by the dashed-line arrow in Fig. 4).

6 Conclusions

In this paper we have investigated how analysis of psychometric properties such as relative item difficulty ordering and point–biserial correlation of adapted MKT items could be used in combination with more qualitative approaches in a high-quality adaptation process. We have shown how the analysis of item difficulty and point–biserial correlation can be used to uncover problematic items. When going back to analyze these items again with qualitative data, new issues of translation and adaptation might be discovered, and new adaptations of items can be made. As a result, the process of piloting adapted items becomes a continuous process of raising the quality of the measures. In addition to this, our study has arrived at a number of findings that have implications for other researchers who attempt to translate, adapt, and use MKT items in other cultural contexts.

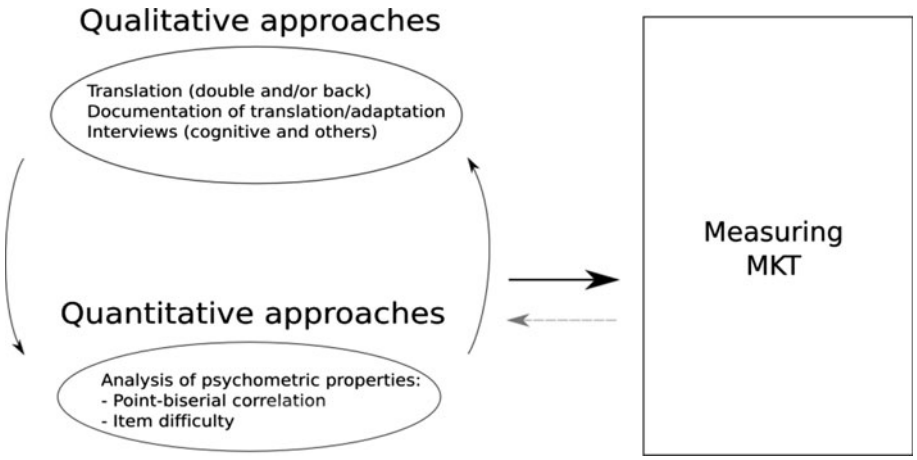

Fig. 4 High-quality adaptation process

Building on the efforts of Delaney and colleagues (2008), we documented and analyzed the process of translating the MKT items into Norwegian (Mosvold et al. 2009). These analyses revealed some issues that appeared similar to those found in other countries such as Ireland (Delaney et al. 2008) and Indonesia (Ng 2011), but we also found language-related and culture-related issues that appeared to be specific to the Norwegian context. Researchers who attempt to translate the MKT items into other languages should pay particular attention to potential local differences in the understanding of seemingly similar terms, such as, for example, "rules". Such differences in understanding have implications for how particular items are understood and interpreted. The MKT items are based on analyses of teaching practice, which might be different across cultures, and researchers who attempt to translate and adapt such items should pay particular attention to such issues.

Other studies have documented and analyzed the translation of MKT items (e.g., Delaney et al. 2008; Ng et al. 2012), interviewed teachers about the items (e.g., Delaney et al. 2008), and discussed the adapted items on the basis of analyses of psychometric properties (e.g., Delaney et al. 2008; Kwon 2009; Ng 2011). In this paper, we have gone one step further and shown how comparison of ordering of item difficulty and point–biserial correlation of items can be used to illuminate problematic items and initiate a new cycle of adaptation. After having uncovered such problematic items from the analysis of these psychometric properties, we went back to conduct a new analysis of qualitative data from the process of translating and adapting the items. This new cycle of analysis revealed translation problems that had not been identified in our previous studies. This indicates that researchers who want to adapt MKT items for use in different cultural contexts need to use different approaches and methods to analyze all parts of the adaptation process, and that the adaptation process needs to become iterative. Evaluation and review should be part of the process, as Fig. 4 indicates.

Analyses of data in our study also indicate that Norwegian teachers seem to understand certain MKT items in ways that differ from those of their US counterparts. Such differences in understanding could potentially have implications for the interpretation of the results, but more research is needed to investigate these possible cultural differences. Researchers who attempt to use the MKT items in other countries should pay particular attention to such differences in teachers' understanding.

When using these different methodological approaches in a cycle of measure adaptation (Fig. 4), researchers can secure higher-quality adapted MKT measures. Such studies also have the potential to add to the existing knowledge of possible cultural differences in MKT.

Acknowledgments Our study is supported by the Norwegian Oil Industry Association, and it would not have been possible without the help of the participating teachers. We would also like to acknowledge the assistance of Professor Michael C. Edwards, The Ohio State University, who provided helpful feedback on the psychometric aspects of the article, and the editor and anonymous reviewers for their useful comments and suggestions.

References

Adams, R. (2005). *PISA 2003 technical report*. Paris: Organization for Economic Co-operation and Development.

Alexander, P. A. (1992). A cognitive perspective on mathematics: Issues of perception, instruction, and assessment. In J. P. Ponte, J. F. Matos, J. M. Matos, & D. Fernandes (Eds.), *Proceeding of the NATO advanced research workshop on advances in mathematical problem solving research. Viana do Castelo, Portugal, 27–30 April, 1991. Research in context of practice, NATO ASI series, series F: computer and systems sciences* (Vol. 89, pp. 61–75). Berlin: Springer.

Anderson-Levitt, K. M. (2002). Teaching culture as national and transnational: a response to teachers' work. *Educational Researcher, 31*(3), 19–21.

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: Springer.

Askew, M. (2008). Mathematical discipline knowledge requirements for prospective primary teachers, and the structure and teaching approaches of programs designed to develop that knowledge. In P. Sullivan, & T. Wood (Eds.), *Knowledge and beliefs in mathematics teaching and teaching development* (pp. 13–35). Rotterdam, Netherlands: Sense Publishers.

Ball, D. L., & Hill, H. C. (2008). *Mathematical knowledge for teaching (MKT) measures. Mathematics released items 2008*. Ann Arbor: University of Michigan.

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). New York, NY: Macmillan.

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.

Bryman, A. (2004). *Social research methods* (2nd ed.). Oxford: Oxford University Press.

Cole, Y. A. (2011). *Mathematical knowledge for teaching: Exploring its transferability and measurement in Ghana*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Darling-Hammond, L. (2000). Teacher quality and student achievement: a review of state policy evidence. *Educational Policy Analysis Archives*, *8*(1). http://epaa.asu.edu/ojs/article/view/392/515. Accessed 27 March 2012.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

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

Edwards, M. C. (2009). An introduction to item response theory using the need for cognition scale. *Social and Personality Psychology Compass, 3*(4), 507–529.

Empson, S. B., & Junk, D. L. (2004). Teachers' knowledge of children's mathematics after implementing a student-centered curriculum. *Journal of Mathematics Teacher Education, 7*(2), 121–144.

Fauskanger, J., & Mosvold, R. (2010). Undervisningskunnskap i matematikk: Tilpasning av en amerikansk undersøkelse til norsk, og læreres opplevelse av undersøkelsen [Mathematical knowledge for teaching: adapting a US developed measure to the Norwegian context and how teachers experience the adapted measure]. *Norsk Pedagogisk Tidsskrift, 94*(2), 112–123.

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). New York, NY: McGraw-Hill.

Graeber, A., & Tirosh, D. (2008). Pedagogical content knowledge: Useful concept or elusive notion. In P. Sullivan, & T. Wood (Eds.), *Knowledge and beliefs in mathematics teaching and teaching development* (pp. 117–132). Rotterdam, Netherlands: Sense Publishers.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist, 27*(3), 353–383.

Hiebert, J., & Grouws, D. (2007). The effects of classroom mathematics teaching on students' learning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Greenwich, CT: Information Age Publishing.

Hill, H. C. (2007). *Technical report on number and operations content knowledge items—2001–2006*. Ann Arbor: University of Michigan, Learning Mathematics for Teaching Project.

Hill, H. C. (2010). The nature and predictors of elementary teachers' mathematical knowledge for teaching. *Journal for Research in Mathematics Education, 41*(5), 513–545.

Hill, H. C., Ball, D. L., & Schilling, S. (2008a). 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. M., Phelps, G. C., Sleep, L., et al. (2008b). Mathematical knowledge for teaching and the mathematical quality of instruction: an exploratory study. *Cognition and Instruction, 26*(4), 430–511.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematical knowledge for teaching. *The Elementary School Journal, 105*(1), 11–30.

Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing teachers' mathematical knowledge. What knowledge matters and what evidence counts? In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 111–156). Charlotte, NC: Information Age Publishing.

Jakobsen, A., Fauskanger, J., Mosvold, R., & Bjuland, R. (2011). Comparison of item performance in a Norwegian study using US developed mathematical knowledge for teaching measures. In M. Pytlak, T. Rowland, & E. Swoboda (Eds.), *Proceedings of the Seventh Congress of the European Society for Research in Mathematics Education* (pp. 1802–1811). Poland: University of Rzeszów.

Jakobsen, A., Fauskanger, J., Mosvold, R., & Bjuland, R. (in press). Correlations between teachers' MKT in different content areas. In *Proceedings from NORMA11, The sixth Nordic Conference on Mathematics Education*. Iceland: Reykjavik.

Kawanaka, T., Stigler, J. W., & Hiebert, J. (1998). Studying mathematics classrooms in Germany, Japan and the United States: Lessons from TIMSS videotape study. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 86–103). London: Falmer Press.

Kwon, M. (2009). *Validating the adapted mathematical knowledge for teaching (MKT) measures in Korea*. Paper presented at the AERA 2009 Annual Meeting.

Ma, L. (2010). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Anniversary edition. New York, NY: Routledge.

Mosvold, R., Fauskanger, J., Jakobsen, A., & Melhus, K. (2009). Translating test items into Norwegian—without getting lost in translation? *Nordic Studies in Mathematics Education, 14*(4), 9–31.

Ng, D. (2011). Using the MKT measures to reveal Indonesian teachers' mathematical knowledge: challenges and potentials. *ZDM—The International Journal on Mathematics Education*, 1–13. doi:10.1007/s11858-011-0375-9.

Ng, D., Mosvold, R., & Fauskanger, J. (2012). Translating and adapting the mathematical knowledge for teaching (MKT) measures: the cases of Indonesia and Norway. *The Mathematics Enthusiast, 9*(1&2), 149–178.

Schilling, S. G. (2007). The role of psychometric modeling in test validation: an application of multidimensional item response theory. *Measurement: Interdisciplinary Research and Perspectives, 5*(2–3), 93–106.

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

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

Stylianides, A. J., & Delaney, S. (2011). The cultural dimension of teachers' mathematical knowledge. In T. Rowland & K. Ruthven (Eds.), *Mathematical knowledge in teaching* (pp. 179–191). London: Springer.

Sullivan, P., & Wood, B. (Eds.). (2008). *Knowledge and beliefs in mathematics teaching and teaching development*., 1 Rotterdam: Sense Publishers.

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.

Tchoshanov, M. A. (2011). Relationship between teacher knowledge of concepts and connections, teaching practice, and student achievement in middle grades mathematics. *Educational Studies in Mathematics, 76*(2), 141–164.

Törner, G., Rolka, K., Rösken, B., & Sriraman, B. (2010). Understanding a teacher's actions in the classroom by applying Schoenfeld's theory *Teaching-In-Context*: Reflecting on goals and beliefs. In B. Sriraman & L. English (Eds.), *Theories of mathematics education. Seeking new frontiers* (pp. 401–420). Heidelberg: Springer.

Zimowski, M. F., Muraki, E., Islevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]*. Lincolnwood, IL: Scientific Software International, Inc.