In the Learning Mathematics for Teaching (LMT) project, measures were developed in order to gauge growth in teachers’ mathematical knowledge for teaching (MKT) and to learn if and how such knowledge contributes to students’ achievement. This paper documents the results from using adapted U.S. developed measures in a pilot study involving 142 Norwegian teachers. Psychometric analyses were performed on the Norwegian measures, and for some item parameters, we find strong correlation to the item performance characteristics found in the U.S.

INTRODUCTION

The study of mathematics teachers’ knowledge and practice has been an active field of research for several decades (Ponte & Chapman, 2006). Inspired by Shulman’s (1986) work, a group of researchers at the University of Michigan have developed a theory of “mathematical knowledge for teaching” (MKT). Part of this research included development of items used to measure teachers’ MKT and to understand the effect such knowledge has on student achievement. It is also used to study and to compare outcomes of professional development of teachers and to improve teacher education. Research shows that teachers with high MKT score can be positively associated with increased learning by their pupils (Hill, Rowan, & Ball, 2005) and with the mathematical quality of instruction (Hill, Blunk, Charalambous, Lewis, Phelps, Sleep, et al., 2008). While some researchers claim that teaching is a cultural activity (Stigler & Hiebert, 1999), little is known to what extent this also applies to the MKT construct.

The measures under investigation were created on the basis of qualitative studies of mathematics teaching in U.S. classrooms and designed to reflect U.S. teachers’ knowledge about the content taught as well as pedagogical content knowledge (Ball, Thames, & Phelps, 2008). Although the items were never made for use outside of the U.S., there have been several attempts to adapt and apply the MKT measures in other countries (e.g. Delaney, Ball, Hill, Schilling, & Zopf, 2008; Mosvold, Fauskanger, Jakobsen, & Melhus, 2009; Ng, 2009). The challenges of adapting and validating these items for use in Norway relate to issues of translation (Mosvold et al., 2009) as well as issues concerning teachers’ experiences and reflections after having worked individually on the measures (in Norway, see Fauskanger & Mosvold, 2010; Bjuland, Mosvold, &
Fauskanger, in progress). As part of this ongoing research, we want to see how the U.S. developed MKT measures perform in Norway and compare with performance in the U.S. This is important if one wants to rely on research and development of MKT theory done in other countries. In this paper, we discuss how the items performed in a pilot study involving 142 primary and lower secondary school teachers in Norway. Our results are compared to results obtained in the U.S. (Hill, 2007). We focus on psychometric properties of the items by addressing the following research question:

Does the U.S. developed mathematical knowledge for teaching items perform in the same manner in Norway?

THEORETICAL BACKGROUND

In recent years, effective professional teacher development has been studied extensively (e.g. Garet, Porter, Desimone, Birman, & Yoon, 2001). According to these authors, however, literature in this field provides little direct evidence of positive outcomes for the participating teachers and their students. Some promising work has been carried out by researchers at the University of Michigan. Their analyses of teachers’ MKT demonstrate that teachers’ MKT made a difference in teachers’ mathematical quality of instruction (Hill et al., 2008) and in pupil’s achievement in mathematics (Hill, Rowan, & Ball, 2005).

Theoretically, the MKT construct follows Shulman’s (1986) work and the categorization of the various components of teacher knowledge that has evolved from Shulman’s original proposal. The work done at the University of Michigan resulted in the model of MKT presented in Figure 1, a model still under development.

The MKT items were developed based on studies of videos from classroom practice, and the domains have been identified both in the U.S. (Ball et al., 2008) and in Norway, where Drageset (2009) has verified the existence of the constructs specialized content knowledge (SCK) and common content knowledge (CCK).

Although the items focus on important tasks of teaching (e.g. presenting mathematical ideas), which are supposed to be of a universal nature (Ball et al., 2008) they may not perform as intended in other countries (e.g. Delaney et al., 2008). This indicates that the translation and adaptation of the MKT measures into a different language (and culture) is not straightforward and requires careful scrutiny and different methodological approaches in order to be successful (e.g. Mosvold et al., 2009).

Investigating what in-service mathematics teachers know is uncommon in Norway, and according to Lysne (2006), assessment in education is a controversial issue in many western countries. In addition, the multiple-choice format is seldom used in Norway, but this seems to be changing (Sirnes, 2005).
Reutzel and his colleagues (in press) claim that the measurement of practicing teachers’ knowledge is not widely accepted, and other assessments than written are highlighted (Baxter & Lederman, 1999).

![Figure 1: Domains of Mathematical Knowledge for Teaching (Ball et al., 2008, p. 403. The domains are defined and discussed in the same reference).](image)

We recognize that no assessment is perfect, and all measurement instruments have their advantages as well as disadvantages. The MKT is no exception (e.g. Kane, 2007; Schoenfeld, 2007). It is clear from the writings of Hill and her colleagues (Hill, Sleep, Lewis, & Ball, 2007) that the goal is to move the debate on assessment of teachers from a debate of argument and opinion to one of professional responsibility and evidence. These authors claim that there is a need for assessment instruments that are designed to “investigate what teachers know, and to associate that knowledge with their professional training and their instructional effectiveness” (ibid. p. 112). From this perspective, it is important to develop different approaches to measure teachers’ MKT. The measures in focus in this paper represent one such attempt.

Schilling and Hill (2007) describe their work on validating the MKT measures, and even when building on their work, we need to be aware of the fact that researchers believe that more efforts need to be made concerning the work of validation (Schoenfeld, 2007) in general and more specific the use of psychometric models as the IRT (e.g. Kane, 2007).
METHODS

Efforts have been made to translate and adapt the 2004 elementary form A (MSP_A04, see LMT 2010) of the MKT items into Norwegian (Mosvold et al., 2009). After the translation phase, a pilot study was organized in order to add to the process of validating the translation and adaptation of the MKT items. The overall aim of the study was to investigate whether and how the MKT measures could be used in a Norwegian context (Fauskanger & Mosvold, 2010). The study includes a quantitative part where 142 teachers’ MKT were measured and a qualitative part where a selection of teachers was interviewed in five focus groups. In this paper, we analyze data from the quantitative data only, using item response theory (IRT) models as an approach.

The form that was used consisted of two parts. Part 1 included a total of 61 items (30 item stems). Part 2 consisted of some questions related to the teachers’ gender, teaching experience and background education in mathematics. Part 1 is the focus of attention here. Figure 2 illustrates the nature of the items. This example asks teachers to respond to a mathematical task situated in a teaching context.

11. Students in Mr. Hayes’ class have been working on putting decimals in order. Three students — Andy, Clara, and Keisha — presented 1.1, 12, 48, 102, 31.3, .676 as decimals ordered from least to greatest. What error are these students making? (Mark ONE answer.)

a) They are ignoring place value.
b) They are ignoring the decimal point.
c) They are guessing.
d) They have forgotten their numbers between 0 and 1.
e) They are making all of the above errors.

Figure 2: Example from the set of released items (Ball & Hill, 2008).

Item Response Theory (IRT) models

Since MKT is not directly observable, the MKT items are meant to relate to the construct and can be viewed as one possible operationalization of the construct. Many measurement models could serve as a link to the observed latent world, and item response theory (IRT) is one such model (Edwards, 2009). The LMT project used IRT models to learn more about item performance characteristics, and in order to compare item performance in Norway and the U.S., we have followed the same approach.

A basic idea in IRT is that an observed item response is a function of person properties and item properties (Edwards, 2009). We have used two IRT models
in the analysis of our data. A two-parameter model was used first, making it possible to identify items with high slope and item difficulty. The higher the slope is, the more variability in items responses is attributable to differences in the underlying construct. Item difficulty indicates the point on the ability axis where an individual would have a 50% chance of endorsing a particular item. Item difficulty is reported in standard deviation units, and 0 is the average teacher ability. Items with negative difficulty indicate easier items, whereas items with positive difficulty indicate more difficult items.

Edwards (2009) found that adequate IRT parameter recovery is possible from as few as 200 respondents. However, a general rule is that the bigger the sample size is, the better the estimates get. The quality of the data also determines the number needed for adequate parameter recovery. Since the number of respondents in our sample is lower than 200 (n=142), we also compared the results coming from a one-parameter model. Missing data is not used in parameter estimation in both models.

In addition to the psychometric analysis, we also present the test information curve and the reliability of the assessment. The test information curve shows how much information items provide for the individual teacher along the ability axis.

RESULTS

In this section, the psychometric analyses are presented and discussed in order to determine if the U.S. developed measures perform well in Norway. We have used the program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) for the estimation and testing of item response theory (IRT) models.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>&lt; -2</th>
<th>[-2, -1&gt;</th>
<th>[-1, 0&gt;</th>
<th>[0, 1&gt;</th>
<th>[1, 2&gt;</th>
<th>&gt; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norway</td>
<td>9</td>
<td>14</td>
<td>22</td>
<td>9</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>U.S</td>
<td>5</td>
<td>19</td>
<td>20</td>
<td>12</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1: Numbers of items within different ability intervals in Norway and U.S.

In order to study item performance characteristics in Norway, we first looked at the output from BILOG-MG using a two parameter IRT model. We used the same item names as in the original U.S. form. Data coming from the Norwegian sample show that one item has a negative point-biserial correlation of -0.116 (Item 17c). This indicates that respondents who answered other items correctly would most likely give the wrong answer to this item, which again indicates that the item is not working well in Norway and should be removed. The average difficulty for the Norwegian items was -0.649, and standard deviation 1.267 (average standard error 0.152), whereas in the U.S. the average item difficulty
was -0.573, i.e. slightly higher. In both countries the scale used consists of a distribution of items of difficulty across the ability spectrum (see Table 1).

Figure 3 displays a scatter plot of the Norwegian item difficulties relative to item difficulties found in the U.S. study (LMT 2004). The correlation between the relative item difficulties is relatively strong (0.804 and p-value < 0.0005). Similar strong correlation is also reported in an Indonesian study where two MKT geometry measures were adapted and used (Ng, 2009).

The average Norwegian item slope was 0.688 with standard deviation 0.287 (average standard error 0.152), while the average item slope found in the U.S. was 0.533. We found the correlation between U.S. slopes and the Norwegian slopes to be rather low (0.375). This can be explained by the low sample size.

Items with slopes lower than 0.5 are normally considered as problematic because they do not discriminate between teachers with high mathematical knowledge for teaching and those with lower mathematical knowledge for teaching. We find that the majority of the items have slopes higher than 0.5, both for items in the adapted form used in Norway (51 out of 61 items) and for the original U.S. form (41 out of 61 items). In both countries, only two items had slopes below 0.3.
Each item also has its own information function that is calculated from the item’s parameters, and the peak of the function is a value of the item difficulty. The item slope determines how peaked the test information curve is, and the higher the slope value is, that item will provide more information around this difficulty level. To understand how our test is functioning as a whole, the items information’s function can be combined into a test information function. The test information curve shows how the test measures teachers across the ability levels.

![Test information function](image)

**Figure 4: Test information function (solid line) and standard error (dotted line) curves for the Elementary 2004 A form.**

In Figure 4 we have plotted the test information function for the adapted Norwegian form. 0 is the mean teacher’s ability score. The peak of the test information function is at -0.75, and this test measures best individuals between -2.5 and 2 standard deviations from the mean ability level. The corresponding standard error for this range is below 0.3, which is under a third of a standard deviation. Reliability for the two-parameter IRT model was 0.9145, higher than what was found in the U.S. study.

Due to the relatively small number of respondents in Norway, a one-parameter IRT analysis was also performed. However, since the U.S. item characteristics are coming from a two-parameter model, we will not compare item characteristics coming from two different models (and for slopes it is meaningless). We observed that the correlation between the U.S. difficulty and the Norwegian difficulty was less strong using a one-parameter model (0.775
compared to 0.804). For the one-parameter model the reliability index of the test was 0.899, and maximum information for -0.8750.

CONCLUSIONS

During the translation of the items and in focus group interviews with teachers, we became worried if the measures would function in Norway. Several teachers expressed concern about some of the items, finding them difficult and hard to answer. Some teachers had expected a test with questions similar to what can be found in a textbook for their own students, and being challenged with the understanding of e.g. nonstandard student solutions in a test situation was frustrating (Bjuland, Mosvold, & Fauskanger, in progress). In our analysis of the data, we find that the scale was composed with items with difficulties over a broad ability range (from -3.683 to 2.443) in the same manner as in the original U.S. scale (from -3.734 to 3.454). We also find that there is a strong correlation between the item property characteristics found in Norway and in the U.S. The reliability index is high (0.9145), and from only looking at the item property characteristics, we argue that the adapted measures are working well in Norway and that the measure performs as intended. The scale is performing well for an individual with the ability between -2.5 and 2 standard deviations from the mean ability level. However, this study has several limitations. First, the sample size is low, making the use of a two-parameter model questionable. Second, our study only examines the content knowledge domain, and we would therefore like to do further studies involving the knowledge of content and students (KCS) and the knowledge of content and teaching (KCT) domains. Third, we also believe that observations in classrooms need to be done in order to investigate if the mathematical quality of instructions is linked to MKT measures in the same way as found in the U.S. For items with difference in item performance, we are also looking into the content of the items trying to see if for instance cultural differences can explain this difference.

NOTES

1. Our research project has been supported by OLF, The Norwegian Oil Industry Association.

REFERENCES


(Eds.), *Examining Pedagogical Content Knowledge* (pp. 147-161). Dordrecht: Kluwer Academic Publishers.

Bjuland, R., Mosvold, R., & Fauskanger, J. (in progress). "I think it should be on the level I normally teach!" – How teacher bias might influence the internal validity when measuring MKT.


