Reliability of the Content Knowledge for Teaching-Mathematics Instrument for Pre-service Teachers

Jim Gleason
Department of Mathematics
The University of Alabama
Box 870350
Tuscaloosa, AL 35487-0350
Email: jgleason@as.ua.edu

Abstract

The Content Knowledge for Teaching Mathematics instrument was developed by the Study for Instructional Improvement and Learning Mathematics for Teaching projects at the University of Michigan to measure elementary school and middle school in-service teachers' mathematical knowledge for teaching to assist in the evaluation of professional development programs for mathematics teachers. This instrument is currently in widespread use among colleges and universities for the purpose of evaluating mathematics education programs for prospective elementary and middle school teachers. Since this is an "off-label use of this instrument, this article establishes the reliability of the instrument among this new population of pre-service teachers.

Introduction

One key component of improving the mathematical education of students is to improve the knowledge of their teachers. This knowledge for teaching is complex and includes knowledge about the subject, the students, the curriculum, classroom management, and so forth. In his Presidential Address at the 1985 annual meeting of the American Educational Research Association, Lee Shulman laid out a construct regarding this knowledge needed for teaching (1986). Shulman divided the construct of knowledge for teaching into three major components: "(a) subject matter content knowledge, (b) pedagogical content knowledge, and (c) curricular knowledge" (Shulman, 1986, p. 9).

In the realm of elementary mathematics, this subject matter content knowledge would coincide with what Liping Ma describes as a "profound understanding of fundamental mathematics" (Ma, 1999b). It is "going beyond knowledge of the facts or concepts" and "understanding the structures" of mathematics (Shulman, 1986, p. 9). In particular, "the teacher need not only understand that something is so; the teacher must further understand why it is so, on what grounds its warrant can be asserted, and under what circumstances our belief in its justification can be weakened or even denied" (Shulman, 1986, p. 9).

The construct of subject matter content knowledge for teaching elementary mathematics may be further divided into common content knowledge and specialized content knowledge (Hill, Schilling, & Ball, 2004; Hill, Dean, & Goffney, 2005; Hill, Dean, & Goffney, 2007). Common content knowledge is "knowledge that is common to many disciplines and the public at large," while specialized content knowledge is "knowledge specific to the work of teaching" (Hill et al, 2007, p. 82).

The Content Knowledge for Teaching Mathematics Instrument

Purpose and History: There are currently many programs in the United States focusing on improving the content knowledge for teaching of mathematics of elementary school teachers. The National Science Foundation's Math and Science Partnership program or Department of Education Math and Science Partnership programs sponsor the majority of these programs. With the funds for these programs comes a requirement for evaluation of the programs.

Because of this demand for instruments to measure the growth of teachers' mathematical knowledge for teaching over the course of these professional development programs, the National Science Foundation's Math and Science Partnership program has funded several programs to create such instruments. The Learning Mathematics for Teaching project at The University of Michigan is one such project and they have developed a series of instruments called the Content Knowledge for Teaching Mathematics (CKT-M) instruments.

Since the CKT-M arose in response to a need of large professional development programs, the development group of the CKT-M instrument determined that the instrument must satisfy certain requirements. These included the need to measure large numbers of participants without taking a large amount of time or money; the reliability of the instrument should be such that it could accurately measure the performance of groups, but not individuals; and the instrument must contain linked forms to use as pretests and posttests (Hill & Ball, 2004; Hill et al, 2004; Blunk, Hill, & Phelps, 2005; Hill, 2007a; Hill, 2007b; Hill, 2007c).

In addition to professional development programs, many pre-service teacher programs are also using the CKT-M instrument. Many of these programs have gone through major revisions in the past few years, partially as a result of No Child Left Behind legislation, which increased the number of hours of undergraduate mathematics courses required of elementary teachers. These changes also developed from a report of the Conference Board of the Mathematical Sciences with recommendations about what mathematical courses and topics should be included in undergraduate programs designed for future teachers (2001).

Since the reliability of the CKT-M instrument was established using experienced in-service teachers enrolled in professional development programs (Hill & Ball, 2004; Hill et al, 2004), these reliability information for these instruments is needed for this new distinct demographic or pre-service teachers. This article will explore the reliability of a single published form including each of the three sub-scales corresponding to the content areas of numbers and operations; geometry; and patterns, functions, and algebra.

Methodology

Data Collection and Sample: Over a period of four academic semesters, 424 pre-service teachers enrolled in mathematics courses for elementary teachers at a large university in the southeastern United States served as study subjects. The students enrolled in these courses had already completed a traditional mathematics course, usually college algebra, but had not yet completed many courses in education and had limited exposure to the elementary classroom. Since these mathematics courses are prerequisites for many of the education courses involved in the elementary education major, nearly all of the participants were in their freshman or sophomore year at the university.

The participants completed the survey instrument during a regularly scheduled class time within the first three weeks of classes during four subsequent semesters. They received an

adequate amount of time so that all participants were able to complete the instrument within the class period.

The participants were 97% female and ranged in age from 19-35 with over 95% being under the age of 22. Additionally, 93% described themselves as Caucasian/White, with 5% African American/Black, and the remaining 2% in other categories.

Instrument: The Content Knowledge for Teaching Mathematics (CKTM) instrument consists of multiple-choice questions designed to gain understanding of an individual's knowledge of mathematical content in the three areas of number and operations; geometry; and patterns, functions, and algebra. To have a better idea of the type of items included in these instruments, an example of an item, chosen from the released items, in the area of number and operation is in the appendices. The actual items cannot be shared due to the use agreement for the instrument.

To compare the reliability of the CKT-M instrument between pre-service and in-service teachers, the analysis used a pre-existing form that contained approximately equal number of items from the three content areas of number and operation; geometry; and patterns, functions, and algebra. The choice of the 2004(B) form was because it has undergone several revisions and has a reported three distinct factors corresponding to three major content areas from pre-service mathematics courses, number and operation; geometry; and patterns, functions, and algebra (Hill, Schilling, & Ball, 2004; Hill, Dean, & Goffney, 2007; Schilling, 2007). The only change from the standardized form is the removal of one item due to a typographical error in some of the copies.

Form Reliability Analyses: Following the structure of the original CKTM form, the items were divided into three distinct sub-scales based upon the mathematical subjects of number and operation; geometry; and patterns, functions, and algebra. Each of the sub-scales was then analyzed using a two-parameter item response theory model in MULTILOG (Thissen, 2003) to correspond with the previous analysis of the form using in-service teachers. The analysis included determining how well the item response theory model fit the observed data followed by comparisons between the models generated using pre-service and in-service teachers of the item parameters, the instrument's information and standard error curves, and the marginal reliability for each of the three sub-scales.

The two-parameter item response theory model generated an item difficulty parameter and an item discrimination parameter for each of the items in the three sub-scales. The two parameters for each item generated by the item response theory model are the core of the model and generate all other results from the model including the item characteristic, item information, instrument information, and standard error curves.

Goodness of Fit: In order to verify that the two-parameter model is appropriate for this instrument, with this population, each of the three sub-scales underwent a goodness of fit analysis. This involved a comparison of the model's estimated ability of the subjects with their measured score using a graphical analysis in addition to a correlation.

In addition to testing the ability of the model to estimate an individual's ability level, it is also necessary to verify the ability of the model to estimate participant performance on each item. The item difficulty and discrimination parameters for each item generate an item characteristic curve which estimates how likely individuals at various ability levels are to answer

the item correctly. This item characteristic curve for the *i*-th item is a logarithmic curve given by the equation

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

where θ is a participant's estimated ability level, a_i is the discrimination parameter, and b_i the difficulty parameter of the i-th item.

For each item, an Average Absolute Standardized Residual determined if the item characteristic curve for that item matched the observed percent correct for the subjects at each estimated ability level (Hambleton, 1991). This Average Absolute Standardized Residual was then compared to the item's ability and discrimination parameters to determine which types of items best fit the observed data.

Item Parameters: The item difficulty parameter is the ability level at which half of the subjects answer the question correctly. Subjects whose ability level is below this difficulty parameter are likely to answer the question incorrectly while those whose ability level is above the difficulty parameter are likely to answer the question correctly. Therefore, the item answer difficulty parameters should vary between around two standard deviations above and below the mean for items appropriate for the sampled population. While the BILOG software (Mislevy & Bock, 1997) used in the analysis of the data collected from in-service teachers restricts the difficulty parameters to this interval, the MULTILOG software (Thissen, 2003)used in the pre-service analysis does not have such restrictions.

Since the standard deviation for the difficulty parameters generated with pre-service teacher data was as high as 5.24, an independent-measures t-test is unable to measure the difference between the parameters generated by the in-service and pre-service teachers. Instead, each item is treated as an individual for a related-samples *t*-test. These parameters were compared for all three sub-scales and the full scale.

The item difficulty parameters are likely different since the pre-service teachers' mathematical knowledge for teaching is similar to, but not as strong as that of the in-service teachers. Therefore, a one-tailed repeated-measures t test was used to measure the significance of this difference.

The item discrimination parameter describes how well an item differentiates subjects at that item's difficulty level. Mathematically, this is the slope of the curve at the ability level equal to the difficulty parameter. Theoretically, the item discrimination parameters should be similar between the pre-service and in-service models, and so a two-tailed repeated-measures t test was used to determine significance.

Instrument Information and Standard Error Curves: For each item, the difficulty and discrimination parameters generate an item information curve from the item characteristic curve $P_i(\theta)$, given by the formula

$$I_i(\theta) = \frac{(P'_i(\theta))^2}{P_i(\theta)(1 - P_i(\theta))}.$$

These item information curves are added together to create the instrument information curve

$$I(\theta) = \sum_i a_i^2 \frac{e^{a_i(\theta - b_i)}}{(1 + e^{a_i(\theta - b_i)})^2}$$

which communicates how much information the instrument provides at various ability levels of the subjects.

The standard error curve is the "standard deviation of the asymptotically normal distribution of the maximum likelihood estimate of ability for a given true value of ability" (Hambleton, 1991, p. 95). These two curves are related in that the instrument information curve is the reciprocal of the square of the standard error curve.

For each of the three sub-scales and the full scale, the instrument information and standard error curves of the model generated using pre-service teacher data was compared to the curves generated using the in-service teacher data. The curves for each of the sub-scales and the full scale were graphed on the same axes to allow for easier comparison even though the ability levels used on the independent axis are different for the two models.

Marginal Reliability: Even though one of the benefits of item response theory is the ability to measure an instrument's reliability for subjects at various ability levels, it is often desired to have a single index of reliability for the entire instrument. Along these lines, one defines a marginal measurement error as

$$\sigma_{em}^2 = \frac{\int_{-\infty}^{\infty} \sigma_e^2(\theta) g(\theta) d\theta}{\int_{-\infty}^{\infty} g(\theta) d\theta}$$

where $\sigma_e(\theta)$ is the standard error function derived from the instrument information curve and $g(\theta)$ is the ability distribution of the sample population.

The marginal reliability (Green, 1984; Thissen, 2001) of the instrument is then defined as

$$\rho = \frac{\text{Variance}(\theta) - \sigma_{em}^2}{\text{Variance}(\theta)}.$$

The marginal reliability of each of the three sub-scales and the full scale were computed using the pre-service teacher data. These reliabilities were then compared to those generated using the in-service teacher data (Blunk, Hill, & Phelps, 2005; Hill, 2007a; Hill, 2007b).

Results

Goodness of Fit: The first method used to verify the goodness of fit for the 2-parameter item response model is to compare the model's estimates of ability to the individual's actual score. For the entire CKT-M scale and the three sub-scales (Number and Operation; Geometry; Patterns, Functions, and Algebra), one can see from Figure 1, there is a perfect fit between the data. This is verified by the correlations between estimated ability and true score being 0.9989 (Full Scale), 0.9984 (Number and Operations), 0.9978 (Geometry), and 0.9978 (Patterns, Functions, and Algebra).

To verify the goodness of fit of the model for the items, the Average Absolute Standardized Residual (AASR) was computed for each item using the item parameters generated using the Full Scale. For the AASR to be meaningful, only ability ranges which include a significant

number of participants is included. Using step sizes of 0.2 in the ability estimates, only the ability range of -0.8 to 1.2 had over 10 participants and was included in the analysis.

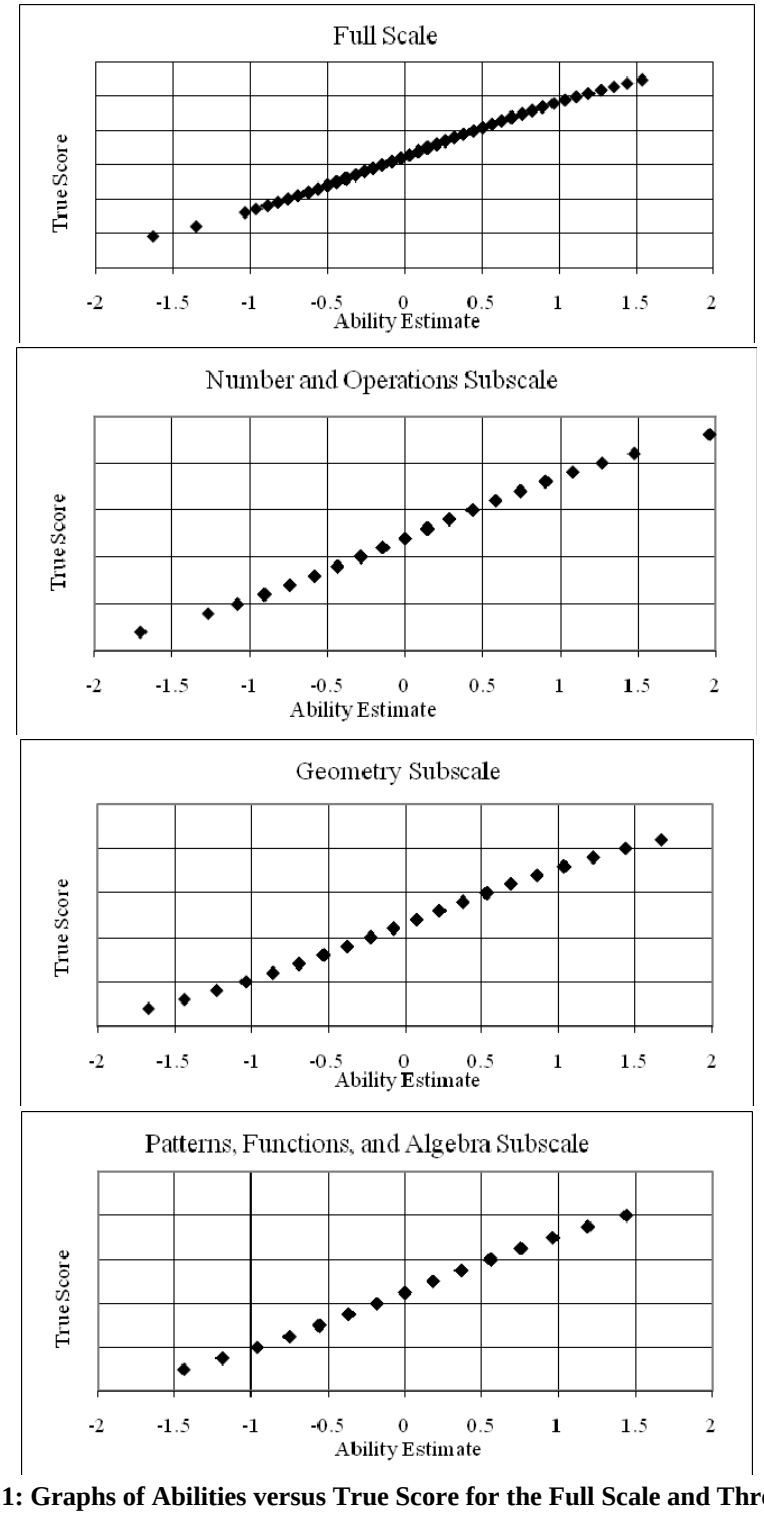

Figure 1: Graphs of Abilities versus True Score for the Full Scale and Three Sub-scales

After removing an outlier with an AASR of 3.73 and difficulty parameter of -541 (almost all subjects answered the item correctly), the AASR's of the items had a mean of 1.20 and standard deviation of 0.34, with a range of 0.63 to 2.31. The correlation of the AASR with the difficulty

parameter was 0.02 for items whose difficulty is within the range of -2.0 to 2.0 and 0.79 for items outside this range. This implies that the items whose difficulty parameter lies outside the range of participants' ability levels do not fit the model well. Similarly, the correlation between the discrimination parameter and the AASR of these items was 0.46 with higher discriminating items not performing as well as lower discriminating items (See Figure 2).

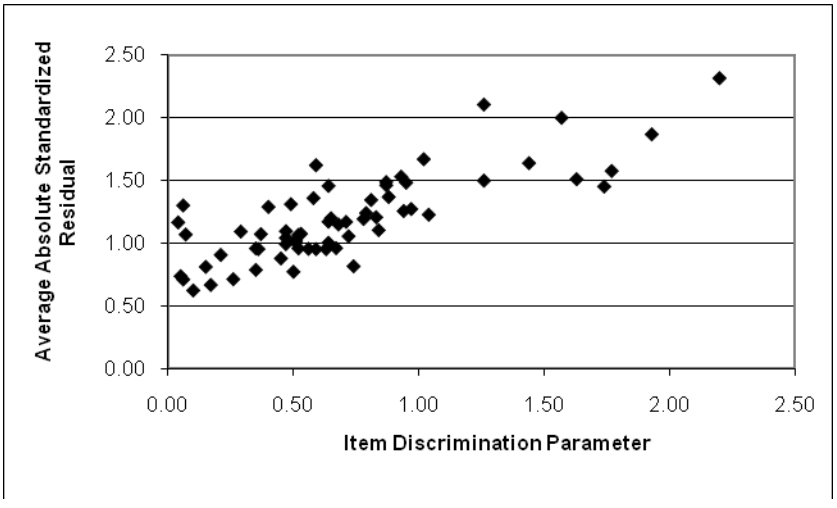

Figure 2: Comparison of the Discrimination Parameter and Average Absolute Standardized Residual of CKT-M Items

Since the 2-parameter item response theory model fits the measured data for the participant ability levels and the item parameters, this model is appropriate for evaluating the usefulness of this instrument with the population of pre-service teachers.

Item Parameters: Since the discrimination parameter is the slope of the item characteristic curve and describes how well an item differentiates between individuals at the item's difficulty level, this parameter should be independent of a population's mathematical knowledge level. As suspected, there is no significant difference between the discrimination parameters generated using in-service and pre-service teachers. The Number and Operation sub-scale ($M = 0.06$, $SD = 0.34$), Geometry sub-scale ($M = -0.04$, $SD = 0.44$), Patterns, Functions, and Algebra sub-scale ($M = -0.09$, $SD = 0.46$), and the Full scale ($M = 0.03$, $SD = 0.41$) all fell within the range on the two-tailed *t*-test for the appropriate degrees of freedom to accept the null-hypothesis that there is no significant difference.

Unlike the discrimination parameter, the difficulty parameter, which measures the point on the ability level where half of the population answers the item correctly, is expected to vary according to the population's overall ability level. For the Number and Operation sub-scale, the difficulty parameters decreased ($M = 1.29$, $SD = 4.68$) between the model using pre-service teachers and the one using in-service teachers. This reduction was statistically significant, $t(23) = -1.35$, $p < 0.05$, one-tailed. Similarly, the Geometry difficulty parameters decreased ($M = 0.95$, $SD = 2.89$) a statistically significant amount, $t(22) = -1.55$, $p < 0.05$, one-tailed. The Patterns, Functions, and Algebra difficulty parameters decreased at an even higher rate ($M = 2.69$, $SD = 6.16$) which was statistically significant, $t(17) = -1.85$, $p < 0.05$, one-tailed. When looking at the

full scale, this decrease of the difficulty parameters (M = −1.56, SD =4.62) was also statistically significant, t(64) = −2.72, p< 0.05, one-tailed.

The decrease in the difficulty parameters fits the hypothesis that the in-service teachers have a significantly higher level of mathematical knowledge for teaching than pre-service teachers.

Instrument Information and Standard Error Curves: The Number and Operation sub-scale exhibited the largest difference between the models using pre-service and in-service teacher data (See Figure 3). For the majority of participants (within one standard deviation of the mean), the instrument provided significantly less information when used with pre-service teachers than with in-service teachers.

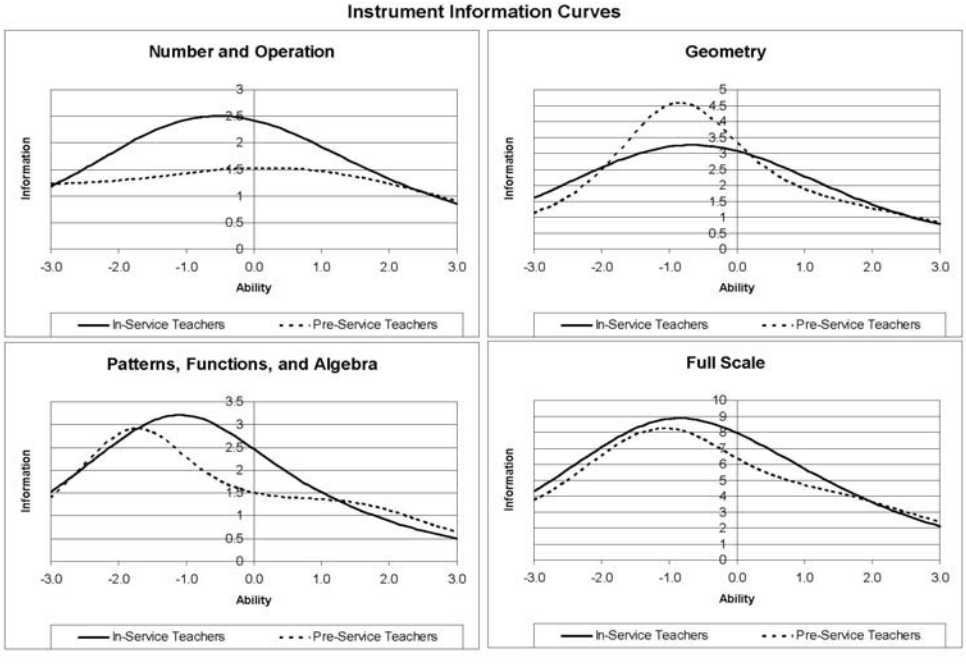

Figure 3: Instrument Information Curves

Furthermore, the standard error for the instrument was always above 0.80 when used with pre-service teachers, while in-service teachers, whose ability ranges between two standard deviations below and one standard deviation above the mean as computed by the model, had a standard error of less than 0.75. (See Figure 4.) This large difference is likely due to the Number and Operation sub-scale measuring the mathematical content most common in the elementary classroom and so the in-service teachers are more likely to have recently worked with the information contained in this instrument.

Since the items on the Geometry sub-scale are focused on subject matter dealt with in the latter elementary grades and middle school, the graphs from this sub-scale show that the instrument likely performed better for pre-service teachers than for in-service teachers since they had seen the material more recently. This phenomenon also occurred with the Patterns, Functions, and Algebra sub-scale to an extent. However, since this sub-scale included questions

regarding exponential growth that almost none of the pre-service teachers answered correctly, the instrument information curve for the pre-service teacher model was lower.

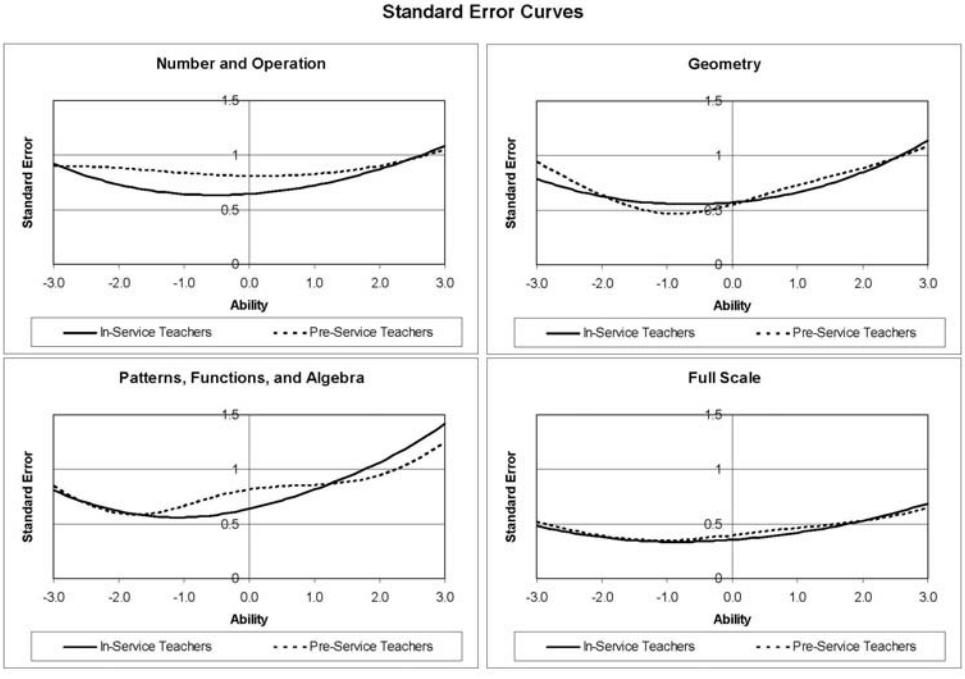

Figure 4: Standard Error Curves

When all three sub-scales are combined into the full scale, both the instrument information curve and the standard error curves for the two models are nearly identical. Furthermore, from these graphs (Figure 3 and Figure 4), one can conclude that the full scale instrument is very reliable with the standard error below 0.5 for nearly all pre-service and in-service teachers.

Marginal Reliability: As with traditional reliability, the marginal reliability is a coefficient between 0 and 1 that measures the proportion of the instrument score is attributed to the actual ability level of the participant rather than noise. For each of the three sub-scales, the marginal reliability is given in Table 1. Since this instrument is designed to differentiate between groups, often as a pre and post test, the reliability indices should be in the range of 0.75 to 0.85 (DeVellis, 1991, p. 85-86). Therefore, the Geometry sub-scale is the only sub-scale near appropriate reliability to use for pre-service teachers.

Table 1: Marginal Reliability for Pre-service and In-service Models

Sub-scale	Pre-service Model	In-service Model
Number and Operations	0.682	0.80
Geometry	0.717	0.861
Patterns, Functions, and Algebra	0.675	0.757

If one combines the three sub-scales to form the full scale, the reliability of the instrument for pre-service teachers becomes 0.8545, which is on the upper end of reliability for use at the group level.

The marginal reliability in each situation above was computed under the assumption that the sub-scales are composed of independent items within the 2-parameter item response theory model. In reality, these sub-scales are composed of several testlets which do not have independence. Therefore, the marginal reliability for the three sub-scales and the full scale are likely significantly lower (Sireci, 1991).

Discussion

Changing the population from in-service teachers to pre-service teachers had a large effect on the item parameters and reliability of the CKT-M form used. The main consequence of this result is researchers should not use the reliability information created using data from in-service teachers when using the CKT-M instrument with pre-service teachers. Researchers should instead make sure that they collect data from enough subjects to run a thorough item response theory analysis on their forms and use these results in reporting their results and should make their own forms specifically for the population of pre-service teachers.

Since the completion of an item response theory analysis is not always possible for every project, there is a need to create specific forms with reliability data generated using pre-service teachers. As evidenced from the results of this study, this form for pre-service teachers may consist of currently developed items from the Learning Mathematics for Teaching item pool, but would not be a previously compiled form. Much work still needs to be completed to determine which items are most appropriate for pre-service teachers and how many items might be needed to have adequate reliability when used with pre-service teachers.

References

Blunk, M., Hill, H. C., and Phelps, G. (2005). Technical report on patterns, functions, and algebra items -2004 Mathematical Knowledge for Teaching (MKT) measures. Technical report, University of Michigan, Ann Arbor, Michigan.

Conference Board of the Mathematical Sciences (2001). *The Mathematical Education of Teachers*, volume 11 of *CBMS Issues in Mathematics Education*. American Mathematical Society, Washington, D.C.

DeVellis, R. F. (1991). *Scale Development: Theory and Applications*, volume 26 of *Applied Social Research Methods Series*. Sage Publications, Inc., Newbury Park, CA.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., and Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21(4):347–360.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*, volume 2 of *Measurement Methods for the Social Sciences*. Sage Publications, Inc., Newbury Park, CA.

Hill, H. C. (2007a). Technical report on geometry items -2002 Mathematical Knowledge for

Teaching (MKT) measures. Technical report, University of Michigan, Ann Arbor, Michigan.

Hill, H. C. (2007b). Technical report on number and operations content knowledge items -2001-2006 Mathematical Knowledge for Teaching (MKT) measures. Technical report, University of Michigan, Ann Arbor, Michigan.

Hill, H. C. (2007c). Validating the MKT measures: Some responses to the commentaries. *Measurement: Interdisciplinary Research and Perspectives*, 5(2-3):209–211.

Hill, H. C. and Ball, D. L. (2004). Learning mathematics for teaching: Results from California's Mathematics Professional Development Institutes. *Journal for Research in Mathematics Education*, 35(5):330–351.

Hill, H. C., Dean, C., and Goffney, I. M. (2005). Assessing "content knowledge for teaching": Data from teachers, non-teachers, and mathematicians. Paper presented at the annual conference of the American Educational Research Association, April 2005, Montreal, Canada.

Hill, H. C., Dean, C., and Goffney, I. M. (2007). Assessing elemental and structural validity: Data from teachers, non-teachers, and mathematicians. *Measurement: Interdisciplinary Research and Perspectives*, 5(2-3):81–92.

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

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Lawrence Erlbaum Associates, Mahwah, NJ.

Mislevy, R. J. and Bock, R. D. (1997). *BILOG: Item analysis and test scoring with binary models [Computer software]*. Scientific Software International, Inc., Lincolnwood, IL.

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.

Sireci, S. G., Thissen, D., and Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3):237–247.

Thissen, D. (2003). *MULTILOG for Windows (version 7.0) [Computer software]*. Scientific Software International, Inc., Mooresville, IN.

Thissen, D. and Wainer, H. (2001). *Test Scoring*. Lawrence Erlbaum Associates, Mahwah, NJ.

Appendix 1: Sample Released Item from CKT-M

Ms. Harris was working with her class on divisibility rules. She told her class that a number is divisible by 4 if and only if the last two digits of the number are divisible by 4. One of her students asked her why the rule for 4 worked. She asked the other students if they could come up with a reason, and several possible reasons were proposed. Which of the following statements comes closest to explaining the reason for the divisibility rule for 4? (Mark ONE answer.)

a) Four is an even number, and odd numbers are not divisible by even numbers.

b) The number 100 is divisible by 4 (and also 1000, 10,000, etc.).

c) Every other even number is divisible by 4, for example, 24 and 28 but not 26.

d) It only works when the sum of the last two digits is an even number.