International Journal of University Teaching and Faculty Development ISSN: 1949-4947
Volume 3, Number 2

EXAMINING THE USE OF THE CONTENT KNOWLEDGE FOR TEACHING MATHEMATICS ITEMS FOR TEACHER EDUCATION PROGRAM ASSESSMENT

Javarro Russell[1]*, Robin Anderson[2], and LouAnn Lovin[3]
[1]National Board of Medical Examiners, Philadelphia, USA
[2]James Madison University, Dept. of Graduate Psychology, Harrisonburg, USA
[3]James Madison University, College of Education, Harrisonburg, USA

ABSTRACT

Undergraduate teacher education programs have sought to identify assessments that can provide a valid indication of the effectiveness of their mathematics education curriculum. This study examines the use of the Content Knowledge for Teaching Mathematics (CKT-M) items for such purpose. Previous research has sought to validate the use of these items for assessing inservice teachers' growth in mathematical knowledge for teaching (MKT). Few research studies have provided validity evidence for using forms of these items to assess the effectiveness of an undergraduate mathematics education curriculum. To obtain validity evidence, this study seeks to confirm the CKT-M's factor structure using data from a sample of preservice teachers (n=988). Results indicate that the use of the CKT-M items may not be appropriate for assessing MKT in preservice teacher samples.

Keywords: Pedagogical Content knowledge; Mathematical Knowledge for Teaching; Preservice Teacher Education; Program assessment

INTRODUCTION

Most teacher educators agree that effective mathematics teachers have a more connected and conceptual understanding of the mathematics they teach than the average person. Lee Shulman (1986) used the phrase "pedagogical content knowledge" to qualitatively differentiate the knowledge held by teachers and the knowledge held by the average person. Liping Ma (1999) used the term "profound understanding of fundamental mathematics (PUFM)" to describe knowledge needed for teaching elementary mathematics for understanding. Deborah Ball used the term "mathematical knowledge for teaching" (MKT) to

* Office: (215) 590-9638 Fax: (215) 590-9440, jrussell@nbme.org

refer to the mathematical knowledge that teachers draw from to teach effectively (e.g., Ball, D., Lubienski, S., and Mewborn, D., 2001; Hill, Schilling, and Ball, 2004). No matter what we call this particular kind of knowledge, there is agreement in the field of education that teachers need a rich understanding of mathematics that is different from what an average person might need. Unfortunately, studies have shown that U.S. inservice and preservice elementary teachers lack the deep understanding of mathematics needed for effective teaching (e.g., Ball, 1990; Borko, Eisenhart, Brown, Underhill, Jones, and Agard, 1992; Eisenhart, Borko, Underhill, Brown, Jones, and Agard,1993; Ma, 1999; National Research Council [NRC], 2001).

Over the past two decades there have been increased calls for accountability in mathematics teacher preparation (Conference Board of Mathematics Sciences [CBMS], 2001; NRC, 2001; American Association of Colleges for Teacher Education [AACTE], 2003; NMAP, 2008). The calls for accountability have created interest in assessing mathematical knowledge for teaching (MKT). Educational policy makers, researchers, and teacher educators alike have expressed an interest in assessing MKT to address their concerns regarding mathematics education. Educational policy makers want to assess the quality of math teachers that are entering the classrooms. Educational researchers are concerned with identifying the type of MKT that is most related to student achievement. Teacher educators, like other educational stakeholders, are concerned about the quality of the programs from which teachers graduate.

Teacher educators' interests in assessing MKT are not strictly for accountability purposes. Some teacher preparation programs that seek to obtain reliable information regarding their program's ability to increase this knowledge do so for the purpose of making curriculum improvements for future cohorts. For this reason, some programs have been in search of appropriate measures that will allow valid inferences to be made about the effectiveness of a university based mathematics education curriculum.

Some teacher education programs rely on the results of licensure examinations such as Praxis I and II to provide some indication of their curriculum's impact on preservice teachers' mathematical content knowledge. These Praxis examinations provide information regarding competency in demonstrating basic content knowledge in mathematics. The Praxis provides feedback about examinees' licensure eligibility based on state departments of education guidelines. However, the feedback lacks specific information needed for assessing the effectiveness of the examinees' math education curriculum. Feedback reports from these types of assessments are summative in nature (Terenzini, 1989). The results do not allow one to ascertain the relationship between the licensure scores and specific aspects of an examinee's teacher education program. Therefore, the validity of using PRAXIS scores to determine the effectiveness of a mathematics education curriculum is, at best, inadequate. Consequently, teacher preparation programs cannot rely on these examinations to make informed decisions about the efficacy of their mathematics education curriculums.

Other programs use portfolios to evaluate preservice teachers' growth in MKT. These portfolios contain products created by the preservice teacher throughout his or her coursework (e.g. lesson plans). They have been touted as excellent tools for allowing preservice teachers to reflect on the growth in their ability to teach mathematics (Cáceres et al, 2010). However, portfolios do not lend themselves to large scale, systematic assessment in most cases. Using portfolios typically requires a much more complex assessment process (e.g. the use of multiple raters), especially in the case of examining growth or development over time.

Research has demonstrated the portfolio's systematic ineffectiveness for assessing teacher ability (Lyons, 2000; Swan, 2009).

Due to the shortcomings of the aforementioned methods for assessing MKT, preservice teacher programs are attempting to identify other measures of MKT. The Content Knowledge for Teaching Mathematics (CKT-M) items are being given considerable attention for use in program assessment. The CKT-M items were developed during the Learning Mathematics for Teaching (LMT) project. These measures have been subjected to theoretical and empirical evaluations to address their ability to measure MKT growth as a function of professional development. However, these validity studies have only been conducted with inservice teachers (Hill, Schilling, and Ball, 2004). The CKT-M measures have not been rigorously examined for use in program assessment with the preservice teacher population. The Standards for Educational and Psychological Testing (AERA, APA, and NCME, 1999) note the responsibility of the test user to provide evidence that the inferences made from test results are appropriate. This evidence is necessary when an assessment is being used for purposes divergent of its intended use.

CKT-M Items

The CKT-M items were developed to measure growth in two specific domains of MKT. The first domain is *Knowledge of Content* (CK). CK consists of two types of content knowledge: *common content knowledge* and specialized *content knowledge*. 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, Dean, and Goffney, 2007, p. 82). Unlike common content knowledge, specialized content knowledge consists of an understanding of mathematical concepts that is necessary for teachers to teach for understanding. For example, you would expect an average adult to be able to compute $3/2 \div 1/3$. However, few adults would recognize that the problem could be modeled as repeated subtraction or as fair sharing, depending on the situation. Nor would they likely be able to explain why it makes sense to flip the bottom fraction and multiply to find an answer. This kind of knowledge is representative of the specialized knowledge needed by teachers. Specialized content knowledge also includes presenting the same concept in different ways, interpreting and understanding different methods of solving problems, evaluating and refining how a textbook approaches particular topics, and providing learners with examples of mathematical concepts.

The second domain is *Knowledge of Student and Content* (KSC). This domain consists of the ability to identify common mistakes students make or common misconceptions they may develop. KSC also includes the ability to identify which representations of mathematical concepts are more likely to help students understand the content. The conceptualization of these knowledge domains was generated through an integration of theories produced by researchers Ball and Bass (2000), Grossman (1990), and Shulman (1987). An in-depth analysis of the nomological network comprising these researchers' conceptualizations of pedagogical knowledge can be found in Hill, et al. (2004).

Prior to writing items, Hill et al. (2004) analyzed student work and reviewed qualitative data on teacher experiences. They also reviewed curricula for the various mathematics content areas. The items were then written to address three mathematical content areas:

numbers and operations; *geometry*, and *patterns, functions; and algebra*. These content areas were chosen because they represent a large portion of the mathematical content taught in kindergarten through sixth grade. The developers ensured that items were devoid of references to any pedagogical technique that would advantage one teacher over another. This prevented the authors from confounding their inferences about teachers' mathematical knowledge for teaching with teachers' mastery of pedagogical technique. These guidelines formed their model for item development.

The psychometric properties of the CKT-M items were initially analyzed through factor analysis (Hill et al. 2004). Factor analysis is a statistical procedure that can identify patterns or relationships among items when developing measures (Gorsuch, 1983). The authors used this statistical technique to build evidence for the existence of the knowledge domains purported to be measured by the CKT-M items. They piloted three forms of selected response items written to represent the aforementioned knowledge domains and all content areas excluding geometry. An exploratory factor analysis (EFA) uncovered factors that represent relationships in the response to sets of items. Results indicated that more than one factor contributed to the responses to several items. However, the researchers determined that the relationships amongst the items were strong in each domain for which they were written. They labeled the domains *knowledge of content in numbers and operations, knowledge of student and content in numbers and operations,* and *knowledge of content in patterns function and algebra*.

In the same study, the authors also conducted a bi-factor analysis in which they found that a substantial number of items loaded onto a general math knowledge factor, as well as a content specific factor. This bi-factor analysis allows researchers to model item responses that are a function of a primary factor, as well as a one other secondary factor. In this case, all items were closely related to the primary factor, general math ability. Variances in item responses were also related to a specific secondary factor, the math content area for which the items were created. However, some items were just as likely to be related to the general math ability as they were related to the secondary factor. Though some evidence of the hypothesized knowledge domains could be inferred from the data, the authors concluded that more studies should be conducted with more items representing the math content areas and two knowledge domains.

Hill et al. (2004) further investigated the properties of each item using item response theory (IRT) methods. This method was used to identify whether multiple forms could be created to measure MKT with adequate levels of reliability. Each form included numbers and operations items, and patterns, functions, and algebra items, written in the CK and KSC domains. The developers found adequate reliabilities across all forms. However, these reliability estimates were likely to be inflated due to the variation in scores on some items being influenced by more than one factor, or trait. For further information on how IRT reliability is impacted by multiple factors or traits, see Sireci, Thissen, and Wainer (1991).

In another study, Hill and Ball (2004) used the items from the numbers and operations scale to assess the development of content knowledge of inservice teachers participating for 1 to 3 weeks in California's Mathematics Professional Development Institutes. Using three parallel forms they obtained pre-test and posttest data on 398 teachers. Reliabilities for these forms ranged from .71 to .78. The authors were able to identify significant growth in MKT scores within their sample of teachers. They also found that the duration of the development institute contributed to growth amongst the teachers. As expected, teachers participating for

three weeks showed more growth in numbers and operations content knowledge than those who participated for institutes lasting less than three weeks. These findings appear to provide support for the possible usefulness of the instrument in the assessment of content knowledge of inservice teachers participating in professional development activities. However, this usefulness is limited to broadly describing MKT growth. The identified growth could not be used to improve particular portions of the professional development programs, nor could they be used to pinpoint particular domains of MKT that exhibited the most growth.

In their 2005 study, Hill, Rowan, and Ball attempted to link teacher content knowledge to student achievement in mathematics. They chose a sample of 334 first grade teachers and 365 third grade teachers from 115 schools participating in a comprehensive school reform program. From the classrooms of participating teachers, the study obtained a first grade cohort of 1330 students and a third grade cohort of 1773 students. The developers administered several instruments to gain information on student achievement, student demographics, teacher background, and classroom characteristics. *Numbers and operations* and *algebra* MKT scales were used to capture content knowledge for teaching mathematics. Using IRT methods for analyses, a reliability estimate of .88 was found for the mathematics knowledge items. Through the use of linear mixture modeling analyses (see Magidson and Vermunt, 2002 for information on mixture modeling) the authors also found that mathematical knowledge for teaching, as measured by the MKT scales, predicted mean student achievement in the first and third grades.

These studies provide some evidence of the CKT-M items' ability to broadly describe growth in content knowledge as a function of professional development. The LMT researchers have also provided some evidence that growth in MKT is related to knowledge domains beyond general knowledge of mathematics. However, there is a lack of evidence that clearly delineates the theorized knowledge domains they suggest contributed to their results. Also, the researchers did not note what aspects of the professional development contributed to the growth in MKT. Consequently, programs using these measures will find it difficult to use these results to make programmatic changes that will contribute to the growth of teachers' MKT.

Despite the studies conducted with inservice teachers, little is known about whether these items can meet the purposes of preservice teacher program assessment. No research has supported that the use of these items could allow teacher education programs to make valid inferences about their effectiveness in developing preservice MKT. To this end, the purpose of this study was to provide validity evidence for the use of this instrument with an undergraduate teacher education population.

There are several ways to address the validity of inferences made from these items when they are administered for preservice program assessment. One way is to conduct a back translation of the items to the curriculum. This process requires that the program review the CKT-M items to determine whether the items' content is being taught within the program's mathematics education curriculum. The extent to which the construct being measured by the items overlaps with the mathematics education curriculum has specific implications for the appropriateness of using the items. The developers of the CKT-M items did not create items to focus on any set curriculum. In fact, their caveats for use indicate that the inferences made from these items could be adversely impacted when a program's participants are able select their course of study (Learning Mathematics for Teaching Project, n.d.). For this reason, a back translation would be a necessary aspect of obtaining validity evidence for the use of the

CKT-M items in program assessment. Another process for validating the use of a measure includes ensuring that responses to the items relate to one another as expected. This process, known as factor analysis, involves piloting the items with the group of preservice teachers and analyzing the structure of their responses. Previous factor analytic work suggests that the items are multidimensional. This means that multiple knowledge domains are being measured by the items. However, the process used in scoring the CKT-M items suggests that a unidimensional structure of content knowledge for teaching mathematics fits the data. This unidimensional structure suggests that an interpretation of only one domain, as indicated by the total score, is appropriate. Determining the most appropriate interpretation of the scores in a preservice teacher population is essential to the process of program assessment.

A confirmatory factor analytic study is used here to test the hypothesized structure of the CKT-M items (Hill, et al., 2004) with a sample of preservice teachers. This procedure allows us to determine the interrelationships among the CKT-M items when used with a sample of preservice teachers. The goal is to determine whether the interrelationships amongst the items are similar to what is observed in an in-service population. If the structure of the items differs between the two populations, then the inferences made about each population differ as well. This would be the major reason for cautioning against using these items for program assessment in an undergraduate teacher education program.

Three different factor structures were applied to the items based on previous research (Hill, et al., 2004). A one-factor (Figure 1) and a three-factor (Figure 2) model were tested first. Adequate model fit for the one factor model would provide support for computing total scores for this sample of students. Adequate fit of a three-factor model would provide support for using the content based sub-scales to aide in the interpretation of MKT. A four-factor bi-factor model (Figure 3) was also tested to provide an evaluation of malformation that occurs when unidimensional models are fit to multidimensional data. The bi-factor model allows items to load onto a general factor and a content specific factor (Gibbons and Hedeker, 1992). This model addresses previous research suggesting MKT can be separated into factors that address general math ability and knowledge specific to the math content being assessed (Hill et al., 2004). Successful fit of this model would suggest that we are able to calculate statistically meaningful total scores, as well as content related scores (i.e. numbers and operations).

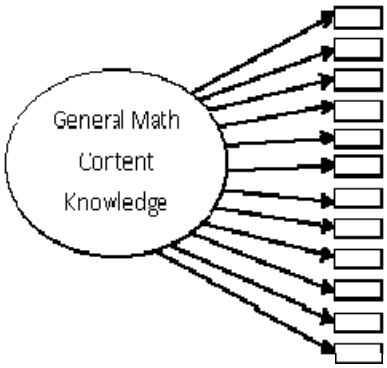

Figure 1. One-factor Model of General Math Content Knowledge.

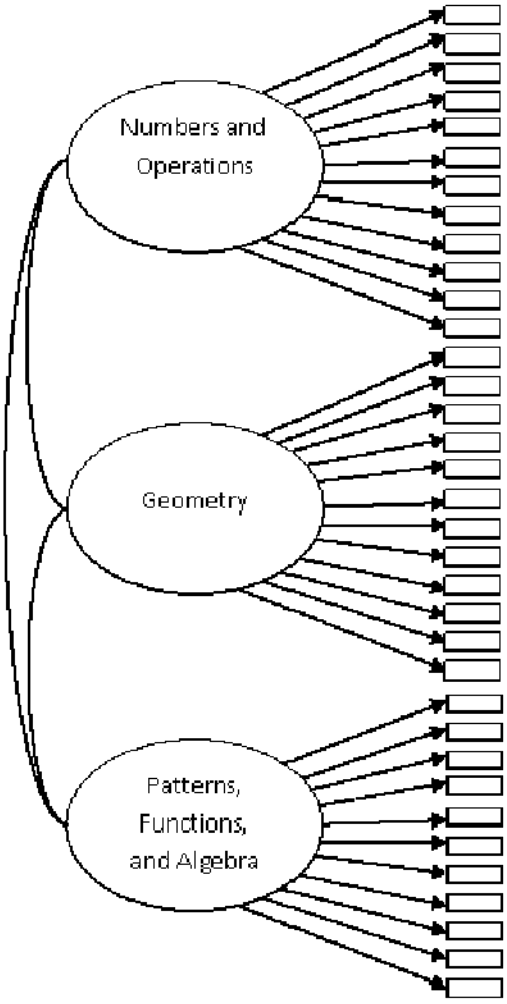

Figure 2. Three-factor model of Math Content Knowledge.

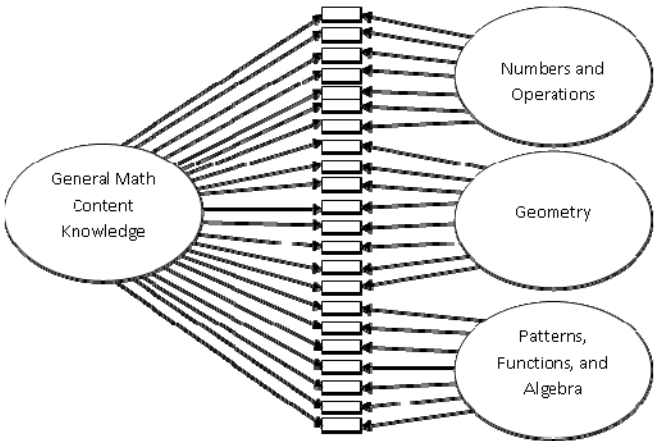

Figure 3. Bi-factor model of Math Content Knowledge.

METHOD

Subjects

To test the validity of the CKT-M items the researchers used archived data from 1013 preservice teachers attending a rural Mid-Atlantic university. Data were collected from past administrations of the CKT-M items spanning fall 2005 through spring 2009 semesters. The items were typically administered within the first weeks of the semester in which the preservice teachers began their first math course. The sample was composed of 90% non-Hispanic white students. Females comprised 96% of the sample. Due to the removal of outliers and cases without responses to all 62 items, a final sample of 988 was used.

Measure

Form A of the CKT-M items (Hill et al., 2004) was administered to the sample described above. There were 62 items in total. Some items exist as part of testlets. Testlets are a group of items that correspond to the same stem. In this study, the testlet items were scored as independent items, instead of subgroups of items. A total score was created for each test taker by summing all correct items. Likewise, sub-scale scores were calculated by summing correct responses for all items identified as representing one of the three content areas.

RESULTS

Data Screening

The data were screened for univariate and multivariate outliers using the Decarlo (1997) macro written for SPSS. There were no univariate outliers. Nine multivariate outliers were identified. Further review of the response pattern for each outlier indicated that each participant answered several items at random. These participants were removed from the data set. In addition to screening for outliers, procedures were used to analyze multicollinearity and normality of the data. Univariate multicollinearity was screened for using the bivariate tetrachoric correlation table and the tolerance values obtained in SPSS. Tolerance values of less than .10 are indicative of multicollinearity. Items 32 and 33 were correlated $r = .814$. Items 17 and 18 were correlated $r = .744$. Items 32 and 36 were correlated $r = .749$. Items 33 and 36 were correlated $r = .737$. Though these correlations were relatively high, the tolerance value obtained for this sample was greater than .10. This indicates no multicollinearity. Univariate and multivariate normality was also evaluated. Skewness and kurtosis values for each item were less than an absolute value of 3 for skewness and less than an absolute value of 8 for kurtosis. This was indicative of univariate normality. Multivariate normality was assessed using Mardia's normalized multivariate kurtosis value. Results indicated that the data was multivariate non-normal.[1]

[1] The correlation matrix was too large to illustrate in this publication. Data requests can be addressed by the authors.

Confirmatory Factor Analysis

The confirmatory factor analyses (CFA) were conducted using Mplus (Muthén and Muthén, 2007). Due to the binary nature of the data, Mplus employed a tetrachoric correlation matrix with robust weighted least square estimation for the CFA (Muthén, 2009). Along with χ^2, model fit was examined using CFI, TLI, RMSEA, and WRMR fit indices. Cut-off values used for assessing the degree of model fit reflect the values suggested by Yu and Muthen (0.96, 0.95,0 .05, and 1.0 respectively; 2002).

Both the one and three factor models fit the data poorly. All indices with the exception of RMSEA failed to meet the cut-off values described above. The model fit data is presented in Table 1. The factor loadings for all items in the single factor model ranged from -.139 to .594. Although, most items had factor loadings below an absolute value of .20. These values, which represent the direct relationship between the factor (i.e. total score) and the item, indicated that there still was a substantial amount of unexplained variance (e.g. variance due to another factor) associated with the items.

Similar results were obtained in the 3-factor solution. Loadings ranged from -.143 to .698, again with many of the loadings small in magnitude. Items 8, 54, and 61 were negatively related to their respective factors, meaning that student who scored highly on the measures scored incorrectly on these items. These items were removed and the one and three factor models were reanalyzed. The change in model fit was minimal and remained unacceptable.

The four-factor bi-factor model would not converge to an admissible solution, and the standard errors of the parameter estimates could not be computed. This is indicative that the model specified did not fit the data.

The residual correlation matrix was examined further in an attempt to diagnose specific areas of misfit. The residuals ranged from -.420 to .507, with most being larger than an absolute value of .350. Identifying a pattern of misfit was difficult due to the large number of relatively high residuals. Therefore, no items were removed due to the size of its residual correlations with other items. Overall, the misfit of these models indicates that a single score or the three sub-scale scores are not appropriate for the data used in the analysis.

Table 1. Fit Statistics for Hypothesized and Modified Models

Model	χ^2	df	WRMR	RMSEA	CFI	TLI
1) 62-item, one-factor	1509.214*	503	1.555	0.045	0.748	0.790
2) 62-item, three-factor	1401.688*	503	1.500	0.043	0.775	0.812
3) 59-item, one-factor	1544.937*	488	1.589	0.047	0.745	0.792
4) 59-item, three-factor	1428.757*	488	1.529	0.044	0.773	0.815

*P= <.001.

DISCUSSION

The CKT-M items are meticulously developed measures of inservice teacher MKT. They are purposed for measuring gains in MKT that occur as a result of professional development. The characteristics of each item directly relate to the purposes for which the measures are to be used. There is also evidence that the structure of the CKT-M items allows for inferences regarding inservice teacher growth in MKT. However, the validity of these inferences falls short on grounds of generalization because the items' structure does not hold when administered to a preservice sample. The validity of the inferences also breaks down in terms of interpretation due to the lack of an empirical link between scores on the CKT-M items and preservice mathematics education curricula.

Using the CKT-M items for program assessment suggests that the effectiveness of a undergraduate mathematics education curriculum can be inferred from preservice teacher scores. However, the belief that the CKT-M items provides insight into program effectiveness is not evidence based.

The items were developed for a purpose that differs from the purpose for which teacher education programs would use the items. Measures used for program assessment that do not capture the intricacy of the education curriculum will contribute little to the information needed to make meaningful inferences about the educational program.

Addressing the structure of the items provides one plausible reason as to why an interpretable score may not be attainable with a preservice sample. Without an adequate model for scoring the items, scores from any CKT-M form will remain ambiguous in interpretation. Thus, any inferences made from these scores will lack validity. Consequently, a rationale for using the CKT-M items for preservice teacher education program assessment is not sustained by the evidence presented here. Validating inferences from assessment scores requires a link between the purpose of the assessment and the use of the assessment scores. This link is obfuscated by an inability to score the data in a meaningful way.

The result of this study is straightforward and succinct. Nonetheless, we acknowledged that most of the preservice teachers in our sample had not completed the entire mathematics education curriculum. The responses to the CKT-M items in this study would only serve as a baseline measurement of preservice knowledge. However, without an interpretable baseline score, inferences regarding growth in MKT cannot be made.

This study suggests that an alternative approach to assessing MKT in a preservice teacher program should be explored. One method is to develop direct assessments of *preservice* MKT. This process would involve defining the scope MKT as it relates to the preservice mathematics education curriculum, developing items that address an empirical definition of MKT, and building test forms that reliably assess growth in MKT at different stages of the curriculum. Forms can then be modified to reflect changes in curriculum that occur as a result of the program assessment.

Future research is needed to develop an assessment process that links a mathematics education curriculum to objective measures of MKT. The field would benefit from studies that provide best practices for objectively assessing the effectiveness mathematics education curriculum.

CONCLUSION

The CKT-M items and the scores derived from the items are designed to address inservice teacher growth in mathematical knowledge for teaching as a result of professional development. They are not designed to address the impact of a mathematics education curriculum on preservice teachers' MKT. Use of these items for purposes such as teacher education program assessment lack empirical support. Inferences made from this or any other alternative use of the items will lack validity. Consequently, any curricular decisions made from these inferences could adversely impact the teacher education program. Therefore, continued use of the CKT-M items for teacher education program assessment is not advisable.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement and Education. (1999). Standards *for educational and psychological testing.* Washington DC: American Psychological Association.

American Association of Colleges for Teacher Education. (2003). *Developing knowledgeable teachers: A framework for standards-based teacher education supported by institutional collaboration.* Washington, DC: AACTE.

Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *Elementary School Journal, 90*(4), 449-466.

Ball, D. L., and Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on teaching and learning mathematics* (pp. 83-104). Westport, CT: Ablex.

Ball, D. L., Lubienski, S., and Mewborn, D. (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: Macmillan.

Borko, H., Eisenhart, M., Brown, C. A., Underhill, R. G., Jones, D., andAgard, P. C. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? *Journal for Research in Mathematics Education, 23,* 194-222.

Caceres, M. J, Chamoso, J. M., andAzcarate, P. (2010). Analysis of the revisions that preservice teachers of Mathematics make of their own project included in their learning portfolio. *Teaching and Teacher Education, 26*(5), 1115-1226.

Conference Board of Mathematical Sciences. (2001). *The mathematical preparation of teachers.* Providence, RI: American Mathematical Society.

Eisenhart, M., Borko, H., Underhill, R., Brown, C., Jones, D., and Agard, P. (1993). Conceptual knowledge falls through the cracks: Complexities of learning to teach mathematics for understanding. *Journal for Research in Mathematics Education,* 24 (1), 8-40.

Gibbons, R. D. and Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika,* 57, 423-436.

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

Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education.* New York: Teachers College Press.

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. (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., Rowan, B., and 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., and Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105,* 11-30.

Johnson, R., Penny, J., and Gordon, B. (2009). *Assessing performance: Developing, scoring, and validating performance tasks.* New York: Guilford.

Learning Mathematics for Teaching (n.d.), Retreived May 15, 2011, from http://sitemaker.umich.edu/lmt/appropriateness

Lyons, N. (1998). Portfolios and their consequences: Developing as a reflective practitioner. In: Lyons, N., Editor, *With portfolio in hand: Validating the new teacher professionalism* (pp. 247–264). New York: Teachers College Press.

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

Magidson, J. and Vermunt, J. K. (2004). Latent class models. In D. Kaplan, Editor, *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Newbury Park, CA: Sage Publications.

Muthén, L. K. (2009, April 13). CFA with binary outcomes. Message posted to http://www.statmodel.com/discussion/messages/9/61.html?1242487452

Muthén, L. K., and Muthén, B. O. (2007). MPLUS (Version 5.1). [Computer software]. Los Angeles, CA: Author.

National Council of Teaching Mathematics (2000). Principles and standards for school mathematics. Reston, VA: Author.

National Research Council. (2001). *Adding it up: Helping children learn mathematics.* J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavorial and Social Sciences and Education. Washington, DC: National Academy Press.

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

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

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*, 1-22.

Sundre, D. L. and Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14* (1), 8-9.

Swan, G. (2009). Information systems in teacher preparation programs: What can we learn from a 5-year longitudinal case study of an electronic portfolio database? *Journal of Educational Computing Research, 41*, 431-451.

Terenzini, P. T. (1989). Assessment with open eyes: Pitfalls in studying student outcomes. *The Journal of Higher Education, 60*(6), 644-664.

Yu, C., and Muthén, B. (2002, April). *Evaluation of the model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.