

Running head: Examination of the LMT

Assessing Mathematical Knowledge for Teaching: An Examination of the Learning Mathematics for  
Teaching Instrument

Javarro A. Russell, Robin Anderson, LouAnn Lovin, Josh Goodman

James Madison University

## Abstract

This study examines the factor structure of the Learning Mathematics for Teaching (LMT) instrument when used for program assessment with undergraduate pre-service teachers. While the one-factor and three-factor models previously identified for in-service teachers did not fit when examined with pre-service teachers, there was reason to believe that a bi-factor model, which specifies all items to load on a general “math ability” factor and subsets of items to load onto a content specific factor, may be more appropriate. However, subsequent analysis of the bi-factor model was not able to converge to a solution.

## Assessing Mathematical Knowledge for Teaching: An Examination of the Learning Mathematics for Teaching Instrument

As a response to increased calls for accountability in teacher training programs, there has been a growing interest in the assessment of knowledge required for teaching mathematics. As the knowledge requirements for teaching evolve, our tools for assessing that knowledge should change as well. One particular instrument currently being used in research studies on the assessment of knowledge for teaching mathematics is the Learning Mathematics for Teaching (LMT) instrument (Hill, Schilling, & Ball, 2004). The knowledge domains the LMT purports to measure have specific implications for teacher training programs. By identifying and developing items to assess the knowledge domains that are important to teaching mathematics, Hill et al. (2004) have provided teacher training programs an opportunity to assess learning outcomes that relate to mathematical knowledge domains. The inferences from these assessments can assist training programs in identifying ways to increase teacher knowledge for teaching mathematics.

The LMT has primarily been used for in-service teacher training program assessment. In-service teachers would typically be administered this instrument during teacher development workshops to assess their growth in mathematical knowledge as a result of those workshops. However, its current use in undergraduate pre-service programs has been sparked by their interest in outcomes assessment related to the mathematical knowledge domains measured by the LMT. Most undergraduate teacher preparation programs are held accountable for their ability to prepare pre-service teachers for the classroom. To this end, programs need to be aware of their students' knowledge, skills, and abilities related to teaching. Program assessment is a means to this end.

### *The Learning Mathematics for Teaching Instrument (LMT)*

The Learning Mathematics for Teaching instrument (LMT) emerged from a project that focused on developing test items measuring the knowledge required for teaching mathematics. Instruments

such as the Praxis already exist to measure mathematical content knowledge; however, Hill et al. (2004) were more interested in developing items that would assist in identifying additional types of knowledge. These types of knowledge are *Knowledge of Content* (CK) and *Knowledge of Student and Content* (KCS). Knowledge of content has two subcategories: *common content knowledge* and *specialized content knowledge*. Common content knowledge consists of mathematic knowledge expectedly held by an average adult. Specialized content knowledge consists of an understanding of mathematical concepts derived from experience in teaching. This includes being able to present the same concept in different ways and understanding different methods of deriving answers to problems. Knowledge of student and content consists of the ability to identify common mistakes students make and how, as well as indentifying students' problem solving strategies. The conceptualization of these knowledge domains was generated through an integration of the pedagogical knowledge theories produced by Ball & Bass (2000), Grossman (1990), and Shulman (1987). An in-depth analysis of the nomological network comprising their conceptualization of pedagogical knowledge can be found in Schilling, Hill, & Ball (2004).

The items for the LMT were written using three math 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 K-6. Hill et al. (2004) began their research for item writing by analyzing student course work, analyzing qualitative data on teacher experiences, and reviewing curricula of the chosen content areas. The developers also ensured the 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 stipulations formed their model for item development. As a result, the authors developed 138 selected response items.

### *Assessing Factor Structure*

Hill et al. (2004) developed and piloted three forms of selected response items written to represent the aforementioned knowledge domains and content areas. An exploratory factor analysis (EFA) was conducted using ORDFAC, a factor analysis software program, to identify the factor structure underlying the measure. This program was chosen for its ability to perform factor analysis on data sets that include subgroups of items corresponding to a common stimulus, called testlets. Although the factors did not align as theorized due to the multidimensional items, the developers concluded that a three factor model was adequate for explaining the data. 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 five-factor bi-factor analysis in which they found that a substantial number of items loaded onto a general math knowledge factor. Items also loaded differentially onto four factors representing items written in a combination of content (e.g. number concepts) and knowledge domains.

Multidimensionality was found, with no firm patterns of loadings. Though some evidence of the knowledge domains was found, the authors concluded that more studies should be conducted with more items representing the three content areas and two knowledge domains.

Hill et al. (2004) further investigated the properties of each item using IRT methods. The items were split into the four areas observed in the previous factor analysis results. A total of 61 numbers and operations CK items, 48 numbers and operations KSC items, 43 geometry items, and 29 patterns, functions and algebra items were piloted on a non-random sample of teachers. The developers found adequate reliabilities across all forms. However, multidimensionality amongst several of the items was apparent. Again, this finding underscores the need for more psychometric work with the instrument.

In another study, Hill and Ball (2004) used the items from the numbers and operations scale to assess the development of content knowledge for in-service teachers participating for 1 to 3 weeks in

California's Mathematics Professional Development Institutes. Through the use of three parallel forms they were able to obtain pre-test and posttest data on 398 teachers. Reliabilities for these forms ranged from .71 to .78. Through the use of mixed models in Hierarchical Linear Models (HLM), the authors were able to identify significant growth in their sample. They also found that the duration of the development institute contributed to more 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 less than three weeks. This further demonstrates the usefulness of the instrument in measuring content knowledge.

In their 2005 study, Hill, Rowan, and Ball attempted to provide validity evidence for the purported link between teacher content knowledge and student achievement in mathematics. They chose a sample of 334 first grade 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. Several instruments were used to gain information on student achievement, student demographics, teacher background, and classroom characteristics. *Numbers and operations* and *algebra* 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 the authors found that mathematical knowledge for teaching, as measured by the LMT items in *numbers and operations* and *algebra*, predicted student achievement in the first and third grades.

#### *Need For Additional Measurement Studies*

Though these studies have provided insight into the instrument's usefulness with in-service teachers, there have been very few inquiries into its use with other samples, such as undergraduate teacher preparation programs (i.e. pre-service teacher development). Recent investments into the use

of the LMT for program evaluation have been made by several undergraduate teacher preparation programs. However, before inferences can be made about pre-service teacher scores on the LMT, some psychometric issues must be resolved.

### Purpose

The responsibility of gathering validity evidence for an instrument rests with the user when the instrument is being used with a population other than the population for which it was created (American Educational Research Association, American Psychological Association, & National Council on Measurement and Education, 1999). To this end, the purpose of this study was to provide validity evidence for the use of this instrument in undergraduate populations. Given the previous work on the LMT, a confirmatory factor analyses was used to examine the factor structure of the LMT instrument. Scores from a student sample were also used. A one-factor (Figure 1) and a three-factor (Figure 2) model were tested to represent the content domains of the LMT (numbers and operations, geometry, and patterns, functions, and algebra). Items representing the knowledge of student and content domain were not included in the analysis due to insufficient item/factor mapping. 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 subscales to aide in the interpretation of the total score. 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 (Chen, West, Sousa, 2009).

It is hypothesized that, a three-factor model representing the content area subscales will fit better than the one- factor model due to the content areas chosen for this instrument. It is also hypothesized that these content areas represent distinct math skills. A bi-factor model is expected to fit

the data better than both the three-factor and one-factor model. Items are expected to load, both onto a general math ability factor and mathematical content specific factors.

## Methods

### *Participants and Procedures*

The current research used archived data from 1013 pre-service teachers attending a Mid-Atlantic university. Data were collected from past administrations spanning fall 2005 through spring 2008 semesters. The test was typically administered within the first weeks of the semester as students began their first math course. The sample was composed of 90% non-Hispanic white students and 96% of the sample was female. Due to the removal of duplicates and cases without responses to all 62 items, a final sample of 988 was used. Data were scanned for outliers using scatter plots in SPSS; no outliers were found.

### *Measures*

A 62 item selected response form of the LMT (Hill et al., 2004) was administered to the sample described above. Testlet items were scored as independent items, instead of subgroups of items. A total score was created for each test taker by summing of all correct items. Likewise, subscale scores (based on the content areas) were calculated by summing correct responses for all items identified as representing one of the content areas. The full scale reliability ranged from .90 to .93 across three previous forms presented in the test manual (Hill, 2004a, 2004b, 2004c). However, this full scale reliability estimate does not include geometry items, or items added to the instrument since piloting in 2002. The range in reliability estimates for the content areas are as follows: geometry, .85 to .92; numbers and operations, .78 to .83; patterns, functions, and algebra, .77 to .80 (Hill, 2004a, 2004b, 2004c). These reliability estimates are for forms created prior to 2004.

Two confirmatory factor analyses (CFA) were conducted using Mplus (Muthén and Muthén, 2007). A single factor solution was fit to the data to verify that the creation of a single total score was

appropriate. Likewise, a three factor solution, where the three factors correspond to the subscale describe above, was fit to the data. 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  $\chi^2$ , model fit was examined using CFI, TLI, RMSEA, and WRMR fit indices. Cut-offs values for assessing the degree of model fit reflect the value suggested by Yu and Muthen (0.96, 0.95, 0.05, and 1.0 respectively; 2002). A bi-factor model was fit to the data using Mplus. This model specified that all items load to a general math factor and items belonging to the three content factors each load to a unique factor. However, even after employing proper estimation methods for dichotomous data, the 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 as it was specified was not correct.

### Results

Both the one and three factor models fit the data poorly. All indices with the expectation of RMSEA failed to meet the cut-off values described above. The model fit data is presented in Table 1. The factor loading of all items in the single factor model ranged from -.139 to .594, with the large majority of them very small in magnitude. Similar patterns were observed in the 3 factor solution. Loadings ranged from -.143 to .698, again with many of the loadings very small. More specifically, items 8, 54, and 61 were negatively related to their respective factors. These items were removed and the one and three factor models were reanalyzed. The change in model fit was minimal and remained unacceptable. Overall, this indicates that a single score or the three sub-scale scores are not appropriate for the data used in the analysis. Model fit was also analyzed using a residual correlation matrix. The residuals ranged from -.420 to .507. Identifying a pattern of misfit was difficult due to the large number of relatively high residuals, though this is indicative of substantial misfit. No items were removed due to the size of its residual correlations with other items.

## Discussion

The aim of this study was to provide validity evidence for the use of the LMT with pre-service populations. In order to determine the validity of the inferences made from scores on the instrument, the dimensionality of the LMT needed to be examined. Each of the models tested in this study have failed to fit the data in terms of both relative and absolute fit indices. Further, an examination of the residual correlations showed many high residual correlations between items, and indicates that misfit was widespread under both models. After the removal of items with negative factor loadings, the fit statistics for the modified models still indicated poor fit. One possible reason for the misfit is the use of testlets in this instrument. Many sets of items were grouped to correspond to a common stimulus. Consequently, subsets of items were correlated due to their dependence on a common stem, instead of solely being dependent on examinee knowledge. This correlation is considered a nuisance trait that causes variance in responses that is unrelated to examinee mathematic knowledge. This causes an increase in the correlation residuals, which adversely affects model fit. However, the removal of items due to the large residuals and negative factor loadings becomes an increasingly complex task when items are grouped into testlets. More investigation into the residuals and the negative factor loadings need to be conducted to provide alternative scoring options, if they are available.

Another reason for model misfit is that the population for which this instrument was created differs from the population used in this study. The items on the instrument may reflect knowledge that pre-service teacher training programs have not provided their students. The difference in classroom experience for the pre-service population and in-service population varies greatly. The pre-service population greatly lacks the experiences that would provide contextual cues for answering some of the items that are focused on interactions with students. Without these cues, the question can become more difficult for this pre-service population. This underscores the responsibility of test users to

conduct validity studies on instruments when they are being used on populations that differ from the intended population.

Another limitation is that the LMT is administered to pre-service teachers in a low-stakes environment. This can cause low motivation for many of the pre-service teachers to put forth effort on a cognitively demanding tests. In low stakes environments, random guessing is likely to occur. Though the instrument provided an “I’m not sure” response option for examinees, no mechanism is provided to determine how the response is used during the test. A better option would be to remove this response option and measure guessing through statistical analysis, if it is of concern.

Considering the findings and the limitations of this study, there are several areas where further research is necessary. First, a model needs to be tested in which testlet effects can be included as a nuisance trait. This would allow for the variance due to dependent items to be modeled. A model such as this would have specific implications for the scoring of the LMT. Also, future studies should quantify pre-service teachers’ levels of training in the area of teaching mathematics. This will allow for an analysis of the impact that different levels of the training program has on mathematical knowledge domains of in-service teachers. Furthermore, teacher training programs need to ensure that they are providing opportunities for students to grow or develop the knowledge domains that are assessed by LMT. If your curriculum objectives do not align to the knowledge domains or content areas of the instrument, then the inferences made from scores on the instrument may not be useful to the program. Lastly, examinee motivation needs to be monitored to ensure that responses to items are not confounded by a lack of motivation (Sundre, 2002).

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement and Education. (1999). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.
- Ball, D. L., & 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.
- Chen, F. F., West, S. G., Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41(2)*, 189-225.
- Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education*. New York: Teachers College Press.
- Hill, H.C. (2004a). Technical report on patterns, functions and algebra items 2001. *Learning Mathematics for Teaching*.
- Hill, H.C. (2004b). Technical report on numbers and operations knowledge of student and content items 2001- 2003. *Learning Mathematics for Teaching*.
- Hill, H.C. (2004c). Technical report on geometry items 2002. *Learning Mathematics for Teaching*.
- Hill, H.C., 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., 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' mathematics knowledge for teaching. *Elementary School Journal, 105*, 11-30.

National Council of Teaching Mathematics (2000). Principles and standards for school mathematics.

Reston, VA: Author.

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

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

Yu, C., & 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.

Table 1.  
*Fit Statistics for Hypothesized and Modified Models*

Model	$\chi^2$	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

Figure 1. One-Factor Model

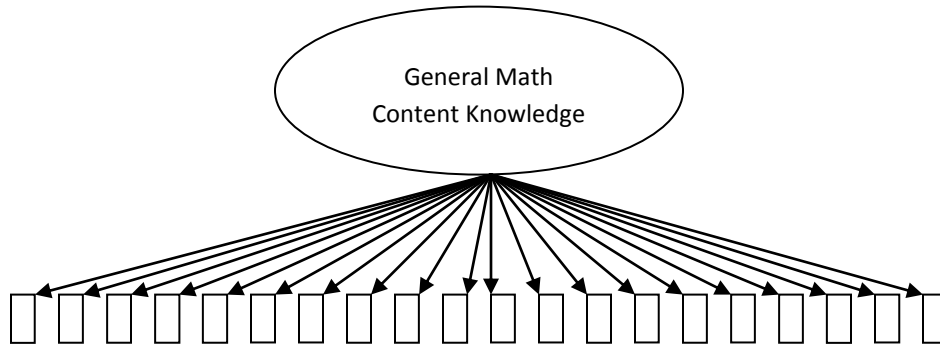
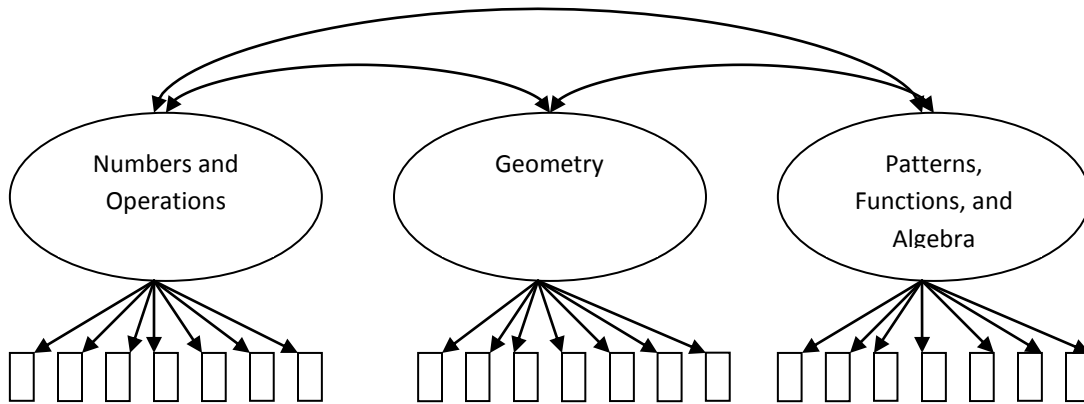
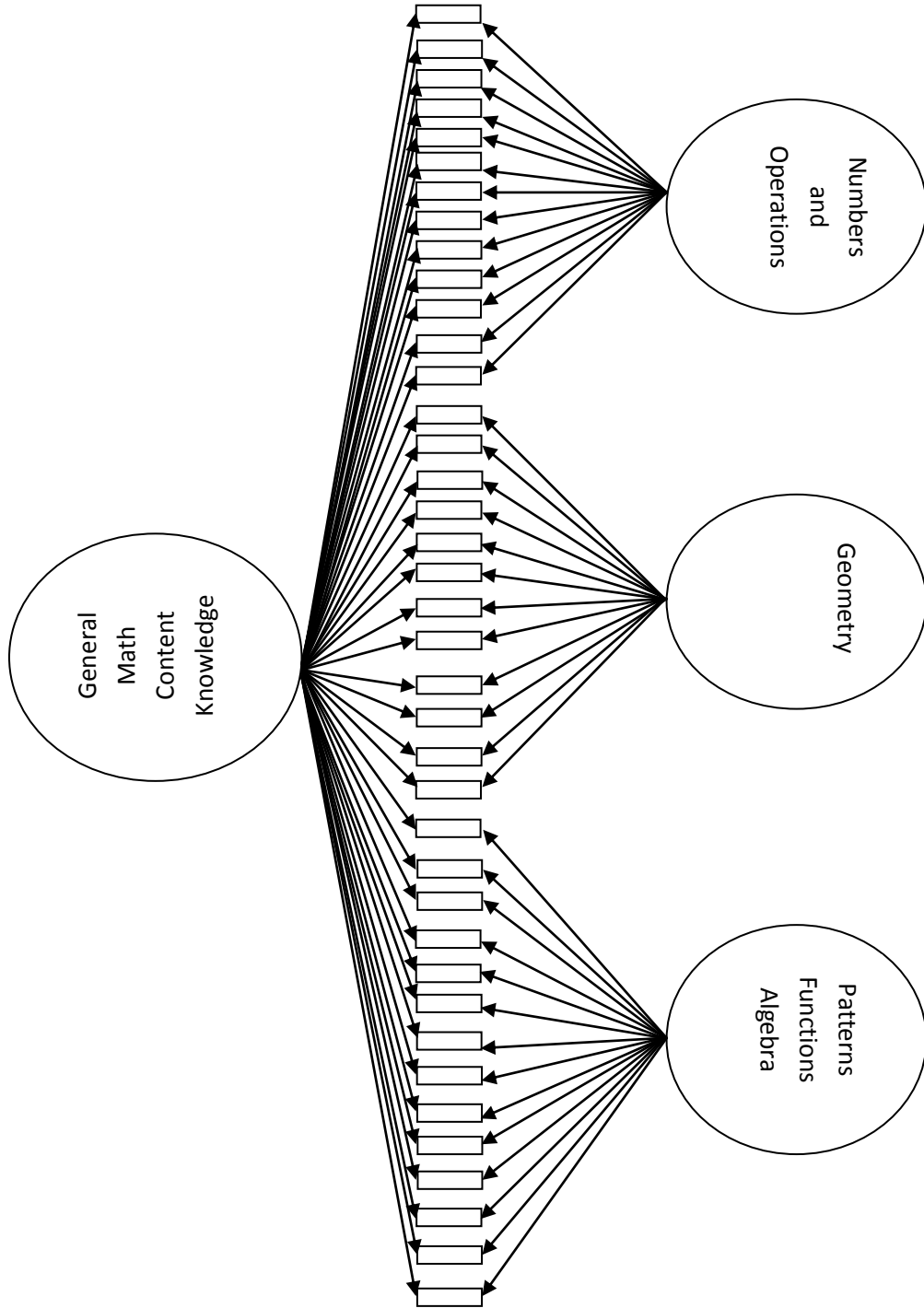


Figure 2. Three-factor Model



\* For illustrative purposes, not all items are shown.

Figure 3. Bi-Factor Model



\*For illustrative purposes, all items not shown.