Spring 2011

Development and validation of the Preservice Mathematical Knowledge for Teaching Items (PMKT): A mixed-methods approach

Javarro Antoine Russell
James Madison University

Development and Validation of the Preservice Mathematical Knowledge for Teaching Items (PMKT): A Mixed-Methods Approach

Javarro A. Russell

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

For the degree of

Doctor of Philosophy

Assessment and Measurement

May 2011

Acknowledgements

I'd like to t express my appreciation for Dr. Robin Anderson, my advisor and committee chair, for being awesome throughout this whole process. The guidance, the motivation, and the free lunch were just enough to get me over the hump. I am thankful for all of the hours she's put into coaching me through the writing process. She is my Coach Dean Smith of Committee Chairpersons.

I'd also like to thank the rest of my committee. LouAnn, thank you for allowing me access to the world of teacher educators. You've helped to provide a lot of insight for the development of this document. Josh, thank you for sticking with me on this project. I truly appreciate your help. Keston, thanks for your words of wisdom and the encouragement.

I'd also like to thank my family for their support. To my parents, thank you for the home cooked meals and the use of your appliances. Thank you for your encouragement. To my wife, thanks for sticking with me throughout this process and for being that light at the end of the tunnel.

Thanks to The Board, The I.C., and RP for those reminders of what awaits me. You all have been running that marathon without me for quite some time. I'm done stretching. Let's go!

Everyone else that has played a part in making this dream a reality, I truly appreciate your support. Most notably, I'd like to thank Dr. Donna Sundre, Dr. Nuria Cuevas, Phil Harris, Dr. Karen White, and Dr. Cara Meixner.

Table of Contents

List of Tables

List of Figures

Abstract

Mathematical knowledge for teaching (MKT) is the knowledge required for teaching mathematics for understanding. Researchers suggest that this construct consists of multiple knowledge domains. Those domains include teachers' knowledge of mathematical content and knowledge about teaching mathematics. These domains of MKT have been theoretically and empirically examined to determine their effects on K-12 student achievement. However, empirical evidence of this relationship is limited due to a lack of measures to assess MKT.

Recently, researchers have constructed measures of MKT to evaluate the effectiveness of professional development activities with in-service teachers. These measures, however, lack validity evidence for use in teacher education program assessment. This process requires adequate tools for assessing the extent to which students meet specific learning outcomes. Previous research has not supported the use of any current measure of MKT for preservice teacher program assessment.

To address this gap in the literature, a process of construct validation was conducted on a scale developed for assessing MKT at the program level of a teacher education program. Validation evidence for the items was obtained using Benson's framework of a strong program of construct validation. The factor structure of the items was analyzed and expected group differences were assessed. Qualitative data from cognitive interviews were then used to provide convergent evidence in regards to the construct validity of the items. The overall purpose of these methods of inquiry was to develop items that would measure the MKT that resulted from a teacher education mathematics curriculum.

Results indicated that an 11- item essentially unidimensional scale of specialized content knowledge could be formed. The factor underlying responses to the scale appeared to be related to specialized content knowledge. Interviews with participants revealed themes indicating that the items were measuring specialized content knowledge. Comparisons among students at differing levels of the mathematics education curriculum revealed significant, but small differences between upper level preservice teachers and preservice teachers whom received no instruction. Further analysis of these items indicated that they could be improved by focusing future item development on examining preservice teachers' misconceptions in evaluating mathematical arguments. Implications of these findings are discussed.

CHAPTER ONE

INTRODUCTION

Mathematical knowledge for teaching (MKT) is a complex construct. It consists of multiple cognitive processes that are often intertwined. For instance, mathematics teachers must demonstrate mastery of the content they intend to teach, as well as demonstrate pedagogical proficiency. Teacher educators and other researchers have been investigating how these cognitive processes develop. Their investigations impact teacher education. Preservice teacher programs consider these researchers' findings in regards to MKT when developing learning outcomes for their mathematics education curriculum. For preservice teacher programs who are concerned about their effectiveness instilling this knowledge, two evaluation-related questions arise: How do we assess the depth of this complex knowledge for teaching? How do we use the results of that assessment to improve teacher preparation programs?

There have been waves of MKT focused research and scale development over the past 20 years (Ball, 1990; Hill, Schilling & Ball, 2004; Darling-Hammond & McLaughlin, 1995; Garet, Porter, Dismone, Birman, &Yoon, 2001; Ma, 1999). The National Council of Teachers of Mathematics (NCTM), a major mathematics educator professional organization, has helped to promote research on mathematics pedagogy by releasing the *Principles & Standards for School Mathematics* (2000). This document guides the improvement of curricula, teaching, and assessment. NCTM specifically notes the need to improve the assessment of teachers, as well as students. As a result of efforts by NCTM and other national teacher organizations, funding for the initiatives outlined in the *Standards* have increased (Behm, 2008). These funds have provided the means for improving preservice mathematical knowledge for teaching. Developing scales to assess MKT at the preservice level is one such way to contribute to this improvement.

The most important component of scale development is the process of obtaining construct validity (DeVellis, 2003). Construct validity refers to the extent to which evidence exists for the inferences made from test scores. Benson (1998) proposed a framework for investigating construct validity. This framework provides direction for navigating the substantive, structural, and external components of a strong program of validation. The substantive component addresses the theoretical and empirical definition of the construct under investigation. The structural component focuses on examining how well the interrelationships amongst the items in a scale represent the definition of the construct. The external stage focuses on testing the relationship between the construct being measured and other constructs. This stage also involves testing hypothesized differences between groups. The research developed herein moves forward with this framework for obtaining construct validity evidence for a scale to measure MKT in a preservice teacher program.

Background

Preservice Teacher Assessment of Mathematical Knowledge

Though lagging behind research on the assessment of in-service teachers, researchers have begun to examine *preservice* mathematical knowledge for teaching. Researchers have focused on preservice mathematical content knowledge (Borko et al., 1992; Gleason, 2010; Graeber, Tirosh, & Glover, 1989), beliefs about mathematics (Kagan, 1992; Richardson, 1996; Scott, 2005), self-efficacy in teaching mathematics (Swars, Hart, Smith, Smith, & Tolar, 2007), classroom performance (Cáceres, Chamoso, & Azcárate, 2010; RPITQ, 2002; Zeichner & Wray, 2001), and ability to determine the achievement of learners (Spitzer, Phelps, Beyers, Johnson, & Sieminski, 2010; Vacc & Bright, 1999). The importance of continued progress in the assessment of preservice teachers is demonstrated in research concerning their deficits in knowledge for teaching mathematics (Ball, 1990; Borko et al., 1992; Leinhardt & Smith, 1985; Simmons et al., 1999; Stacey et al., 2001). These studies broadly suggest that preservice MKT could be improved by making changes to mathematics methods courses and practicum experiences. However, they do not provide suggestions for determining the extent to which these changes should be made.

Guiding this current study are two noteworthy themes from the research. First, there is uncertainty on how best to interpret or operationalize the type of knowledge that is necessary to effectively teach mathematics (Ball & Bass, 2000; Ball, Thames, & Phelps, 2008). Second, there is little mention of how to improve teacher preparation programs based on assessment findings related to MKT. For example, most of the research that has uncovered important findings regarding MKTor personal beliefs about mathematics has been individual focused and qualitative in nature (Hill, Blunk et al., 2008). Researchers have not focused on the use of those findings for program assessment.

MKT and Program Assessment

Program assessment is the process by which inferences are made regarding a curriculum’s effect on student learning. In order to engage in this process, programs must be able to define the outcomes that are expected as a result of the program. Once those outcomes are defined, the programs must identify ways to assess them. This means adopting measures and procedures that will allow the program to make inferences to the extent to which students attain the specified outcomes. Those inferences are then used to inform changes to the program.

Teacher education programs have adopted standards, goals, or outcomes that explicitly address preservice teacher attainment of mathematical knowledge for teaching. The National Council on Accreditation of Teacher Education (NCATE) requires that teacher preparation programs demonstrate effectiveness in assisting preservice teachers in meeting these standards. As a consequence, teacher education programs have been in search of tools that will allow them to systematically assess the mathematical knowledge for teaching gained by their preservice teachers.

Teacher preparation programs typically assess mathematical knowledge for teaching through the use of state licensure examinations (Capraro, Capraro, Kulm, & Raulerson, 2005), portfolios (Frid & Sparrow, 2003; Hartman, 2004; Romberg, 1995), Teacher Work Sample assessments (Girod, 2002), and selected response examinations (Gleason, 2010; Kahn, Cooper, & Bethea, 2003; Mathews & Seamen, 2007; Quinn, 1997; Russell, Goodman, Anderson, & Lovin, 2010). State licensure exams have proficiency standards that are set by each state's licensing board. These exams provide information about which students in each program achieve the proficiency standards. Though this information is important, it does not contribute to a program's understanding of how well their preservice teachers meet the program's learning outcomes. Scores on licensure examinations are unlikely to be informative to a program if the test is not aligned with the program's learning outcomes effectiveness (Nichols & Sugrue, 1999).

Another tool, portfolios, allows preservice teachers to reflect on the growth in their ability to teach mathematics (Cáceres et al, 2010). However, researchers have not demonstrated the systematic effectiveness of portfolios in program assessment (Lyons, 1998; Swan, 2009). Research on the use of Teacher Work Sample (TWS) assessments has shown that these assessments can assist investigating the instructional decisions of preservice teachers. However, like other types of portfolios, TWS have not been demonstrated as an effective tool for assessing mathematical knowledge for teaching at the program level.

Recent attempts to assess mathematical knowledge for teaching include the development and use of selected response measures (Hill et al., 2004; Mathews et al., 2007). Some of these measures were developed specifically for in-service teachers. However, teacher preparation programs have often adopted these measures without fully investigating their psychometric properties with the preservice population (e.g., Swars et al., 2007). Russell et al. (2010) identified several obstacles to making valid inferences from these items. These obstacles include psychometric issues related to item dependency (Serici, Thissen, & Wainer, 1991), the use of an “I don’t know” distractor (Thissen, Steinberg & Fitzpatrick, 1989), and highly difficult items (Reckase, 1985). There are also assessment design issues related to the lack of congruence between the knowledge domains being assessed and the learning outcomes of the preservice teachers (Erwin, 1991). For example, measures that attempt to directly assess mathematical knowledge for teaching (MKT) have been developed using theoretical models that may not represent the actual instructional practice in the preservice teacher program. In fact, the measures have typically been designed with no particular curriculum or program in mind (Hill et al., 2004). Consequently, these measures lack validity evidence for measuring MKT as it is defined by the teacher education program.

Issues in Assessing Preservice MKT at the Program Level

Defining the Construct. The construct, mathematical knowledge for teaching, consists of the required knowledge specific to the job of teaching mathematics. Shulman (1986) was among the first to consider these various knowledge domains and formalize them in writing. What started as knowledge of content, pedagogy, curriculum, and pedagogical content has developed into domains consisting of common content knowledge, specialized content knowledge, knowledge at the mathematical horizon, knowledge of content and student, knowledge of content and teaching, and knowledge of curriculum (Shulman, 1986; Hill, Ball, & Schilling, 2008). These theoretical developments have spawned efforts to create measures to assess this knowledge. However, the relationships among these theorized knowledge domains are complex and have not been thoroughly explored (Ball, Thames, & Phelps, 2009; Kane, 2007). Researchers have suggested that future work continue to operationally define the mathematical knowledge domains (Hill et al., 2004). Indeed, improper specification of the knowledge domains to be assessed affects all subsequent steps of the assessment process.

Mapping MKT to Student Learning Objectives. Prior to assessing MKT in a preservice teacher program, its faculty must identify learning objectives that capture the knowledge they intend to assess (Erwin, 1991). Teacher preparation programs that have aligned with the NCTM *Standards* have, at a minimum, acknowledged that preservice teachers should attain MKT at a level deeper than basic knowledge of mathematics. However, to effectively implement an assessment process for MKT, the program must provide further detail that explains what is meant by a deep level of MKT. The domains of MKT need to be stated in clear and measurable ways. This allows for the assessment of MKT to be more transparent to those who are being assessed, and to those who desire to use the results for program improvement. A lack of clarity in what is being measured reduces the validity of the inferences that can be made from the measurement (Kane, 2007).

Identifying the appropriate measure of MKT. After clearly defining the learning outcomes of the program, a measure must be chosen to assess those outcomes. Preservice teacher programs have an option of selecting an instrument or developing their own. The decision to develop or select a measure will be driven by the pros or cons of either approach. These issues can include cost, measure to objective match, or test properties (Suskie, 2009).

Another issue that may be considered when selecting an appropriate measure is whether to use an objective test or a performance assessment. In teacher education assessment, there is a complex relationship between the factual content knowledge that typically lends itself to objective assessment (e.g., Mathews& Seamen, 2007), and pedagogical content knowledge that is typically measured through performance assessment (e.g., Chapman, 2005). Decisions to use either assessment approach involve examining issues related to assessment's purpose, psychometric properties of the measure, construct representativeness, authenticity, fairness, and costs (Cizek, 1991; Johnson, Penny, & Gordon 2009; Wiggins, 1991). Regardless of the choice, the chosen measure must include an adequate sample of items or tasks that represent the knowledge domain being assessed (Messick, 1989). Additionally, the knowledge domain must align with the learning outcomes that the preservice teachers should achieve as a result of the program.

High stakes versus low stakes assessment. In program assessment, scores are typically used for making decisions concerning the effectiveness of the program instead of decisions about individual students. In some cases, an individual's performance on the assessment does not affect his or her academic standing. Cases such as these are considered low stakes testing environment because consequences are not attached to individual examinee performance. This type of testing environment can have intended as well as unintended effects on the motivation of examinees (Wise &Demars, 2005). The effects of inflated or deflated test taking motivation can introduce bias in test scores, thus adversely impacting score interpretation (Haladyna& Downing, 2004; Sundre &Kitsantas, 2004; Wise et al., 2005). Therefore, techniques involving the removal of scores from low motivated students can be used to address this issue (Wise, Wise, &Bhola, 2006). The use of trained proctors has also been shown to improve motivation during low-stakes testing (Lau, Swerdzewski, Jones, Anderson, &Markle, 2009).

Reconciling Research and Practice

Both qualitative and quantitative research on MKT have provided snapshots of preservice teacher experiences using this type of knowledge (Lowery, 2010; Mathews& Seamen, 2007; Russell et al., 2010). However, few studies have identified appropriate assessments to be used for improving preservice teacher programs. Instruments developed for in-service teachers have not demonstrated adequate psychometric properties when used with preservice teachers (e.g., Russell et al., 2010). The popular use of portfolios has not produced a standardized framework for use in program assessment (e.g., Frid & Sparrow, 2010). Licensure exams have yet to provide feedback appropriate for usein improving preservice teacher programs. Despite these instrument shortcomings, researchers continue to make claims about the development and use of this knowledge in practice. Few studies have empirically examined the relationship between preservice teacher MKT and the learning outcomes of teacher education programs. This exemplifies the lack of synthesis between theory and practice in teacher education (Ball, 2000; Korthagen & Kessels, 1999; Zeichner, 2010). Research conducted on MKT in preservice teachers has provided further understanding of the construct within this population. However, there continues to be a lack of research that demonstrates how the inferences made from measures of MKT can be used to effectively improve preservice teacher education.

Statement of the Problem

The majority of research conducted on MKTwasconducted at an in-service level where most accountability efforts to assess teachers have been focused (Behm, 2008). In many teacher preparation programs, the assessment of MKT is conducted using measures that were not created for program-level inferences (e.g. licensure exams, portfolios, &TWS) or for pre-service populations (e.g. CKT-M items). Consequently, faculty members and program administrators are unable to use the results of these assessments to make informed decisions about curricular changes. Therefore, the development and validation of items that will allow for effective and efficient assessment of preservice MKT is necessary. These items would allow for the program to determine student's strengths and weaknesses. The scale would also assist the program in determining its effectiveness on decreasing students' deficits in MKT.

Purpose

The purpose of this mixed-methods study was to develop the PMKT items for measuring MKT in a preservice sample. The development of these items included a process of construct validation in which the researcher sought to determine how well the items measure MKT. The validity of the inferences that can be made in regards to program assessment was evaluated.The first component of this study involved the administration and analysis of newly developed items for measuring MKT. Students participating at different levels of teacher preparation mathematics courses were administered the items. A comparison group consisting of non-teacher education students was also administered the items. The second study component consisted of qualitative inquiry of pre-service teachers' experiences when responding to items designed to assess their MKT. Think-aloud interview data from pre-service teachers (PTs) in an undergraduate teacher preparation program was collected, transcribed, and analyzed to explore these experiences. The data collected from this explanatory follow-up was used to provide additional validity evidence that suggests the PMKT items are measuring MKT.

Research Questions

The research questions of this study were focused on examiningthe validity of inferences that could be made as a result of using the PMKT items in program assessment. The primary questions were directed toward analyzing the psychometric properties of the items. The secondary questions were focused on obtaining qualitative information regarding the functioning of the items

The following research questions assisted in addressing the purposes of this work:

1. How do the PMKT items perform?
    - What factor structure is plausible for the data?
    - How do item difficulty and item discrimination vary?
    - How effective are the distractors?

  - What level of reliability is demonstrated by these new items when used with this pre-service teacher sample?

2. How do groups of pre-service teachers at different academic levels compare on their aggregate scores?
3. How do pre-service teachers conceptualize (i.e. think about) MKT and what does this conceptualization imply about the development of items for assessing MKT?
4. What level of face validity do the new items created to assess mathematical knowledge for teaching have with a pre-service teacher sample?

Research Design

With the acknowledgement of the philosophical basis for conducting mixed method research, this study attempts to move forward with a pragmatist viewpoint suggesting that the mixing of research designs is a viable way to answer research questions (Creswell, 2003). In this study, qualitative data (qual) was connected to the heavily weighted quantitative data (QUAN) to explore the cognitive process involved in answering the PMKT items.

Research questions one and two attend to the psychometric quality of the items used for measuring MKT in this preservice teacher sample. The need to answer those questions is consistent with the substantive and external stages of construct validation (Benson, 1998). Answering these questions allows the researcher to make informed inferences about the construct under investigation.

Research questions three and four attend to the qualitative nature of the process in which MKT is used at the preservice teacher level. These questions help to explore preservice teachers' conceptualization of MKT. Answering these questions requires interviewing techniques that allow for the investigator to elicit verbal responses from the participants (Merriam, 2009). These techniques allow for a two-way process in which clarification can be sought, thus providing more data to strengthen the validity of inferences made about the construct under investigation (Drennan, 2003).

Limitations

This study attempts to combine the theoretical foundations of two methods of scientific inquiry. In doing so, this study attempts to utilize the strengths of both approaches; however, the use of these combined methods does not preclude the weaknesses of either design from limiting the interpretation of the results. First, the ability to generalize from this study is adversely affected by the samples of participants chosen. The sample size of this study is relatively small and is unlikely to be similar to the national sample of preservice teachers. However, this researcher only hopes to generalize to preservice teachers attending the host university.

Motivation is another possible limitation to this study. An examinee's motivation can affect performance on assessments. This is especially the case when the testing environment is low stakes (Sundre & Kitsantas, 2004), which is the case here. In this study motivation will be examined through the use of an examinee motivation survey (Sundre & Moore, 2002).

Another limitation is related to the treatment. Each level of math course provided by the program has multiple sections. There is the possibility of a lack of standardization across the sections of each course. Students cannot be randomly assigned to courses. Therefore, comparisons among students in different sections of the same course will not be made.

Implications for Research and Practice

The MKT gained as a result of participating in a preservice teacher program is of central importance to the theoretical foundation of the PMKT items. The PMKT items are being constructed to measure MKT that is specific to the learning outcomes of a mathematics education curriculum. By exploring preservice MKT at the program level, this study shifts the framework of measuring MKT. Instead of focusing on a nebulous conceptualization of the construct for descriptive purposes, this researchnarrows the MKT construct domain to the knowledge that is relevant in the teacher preparation programs. Narrowing the domain allows researchers to develop more precise tools for exploring their program's effectiveness in instilling MKT.

There is little research that focuses on assessing mathematical knowledge for teaching specifically for improving preservice teacher programs. Instead many studies focus on describing deficits in this knowledge at the individual student level. These deficit studies do not provide findings that can be used to improve preservice education programs (i.e., curricular programming) based on the measurement of MKT. This research attempts to fill this void by developing items to assess MKT at the preservice program level.

For teacher educators working with preservice teachers, the development of the PMKT items enables an important process for measuring learning outcomes. By addressing the relationship between the complex construct of MKT and the PMKT items, this study engages in *a strong program of construct validation.* This process is intended to help preservice teacher programsmake valid inferences from a measure of the MKTfor a preservice population. This assessment of MKT can also provide indications as to the improvement of the program's mathematics education curriculum.

This study also advances the use of mixed-methods for instrument development. Mixed-methods research continues to be touted as a powerful method of inquiry; however, more information as to the best practices in its use for validation research is necessary (Leahey, 2007). More examples of how to best mix the two paradigms can increase general understanding of this approach (Morell& Tan, 2009).

CHAPTER TWO

LITERATURE REVIEW

Research on mathematical knowledge for teaching (MKT) frames a set of questions that guide inquiry into the knowledge requiredto produce favorable outcomes in K-12 student learning. Researchers and teacher educators are interested in identifying how this knowledge develops, how it relates to other constructs, and how much is necessary to affect changes in K-12 student outcomes. To respond to these inquires, an operational definition of the MKT construct must be presented and a measure must be developed.

The purpose of this study is to engage in the process of developing a measure of MKT for use in a preservice teacher program. To provide a context for this study, this chapter reviews relevant topics related to the measurement of MKT for program assessment. Those topics are: 1) the call for accountability in teacher education and its impact on the research of mathematical knowledge for teaching (MKT), 2) defining MKT in terms of preservice teacher program assessment, 3) strategies for measuring MKT, 4) options for validating measures of MKT.

Call for Accountability in Teacher Education

Near the turn of the century two major professional teacher organizations released reports outlining their expectations of what teachers should know and be able to do. The National Commission on Teaching and American's Future (NCTAF, 1996) provided a much publicized report outlining the need for teacher education programs to increase their standards to ensure the continued improvement of pedagogical quality. Four years later the National Council for Accreditation of Teacher Education (NCATE, 2000) developed the Unit Standards, which set higher requirements for institutions when demonstrating effectiveness in educating future teachers.

NCATE released a report in 2006, *What Makes a Teacher Effective,* articulating the findings of multiple research efforts concerning the quality of teacher preparation. One of the key findings indicated that improvements in teacher education have and will continue to improve the quality of teachers, increase teacher retention, and improve K-12 student learning. The NCATE document hinted that these improvements were related to the increased standards set by their organization. Their research also indicated that nations who invest *heavily* in preservice teacher learning are the leading industrialized nations. They noted the following conclusions and policy recommendations:

> 1) High quality pre-service teacher preparation provides beginning teachers with the knowledge and skills needed for effective teaching in today's heterogeneous classrooms; (2) Programs that circumvent high quality pre-service teacher preparation place the beginning teachers at a disadvantage; (3) High quality pre-service preparation should enjoy strong support from federal, state and local policy; (4) All preparation programs should provide evidence that they prepare candidates with the foundational knowledge and skills to positively affect student learning, or they should be closed--NCATE accredited institutions must provide such evidence; (5) All pathways to teaching should undergo review according to national standards; (6) Professional development schools should become the norm for teacher induction; (7) Many hard-to-staff schools should be re-configured as professional development schools; and (8) More comprehensive assessments of teacher knowledge and performance are needed for teacher licensing (NCATE, 2006, pp. 16-17).

To follow through on the applicable NCATE and NCTAF recommendations, teacher preparation programs needed to focus on the learning outcomes of their candidates and ensure that those outcomes are congruent with the standards and expectations of their professional organization and accrediting bodies. They also needed to ensure that teacher preparation programs provided their candidates with educational activities and experiences that allow them to achieve the learning outcomes and meet the standards set before them. To address educational impact, these programs also needed to develop appropriate assessment strategies to determine the effectiveness of their teacher education curriculum.

In moving forward with the recommendations of the aforementioned national teachers organizations,the National Council of Teachers of Mathematics (2000) and the National Mathematics Advisory Panel (2008) began focusing efforts on understanding what mathematics teachers should know and how that knowledge impacts K-12student learning. The NMAP report noted few studies having empirically investigated the breadth and the impact of teachers' mathematical knowledge. In fact,this lack of empirical evidence existed due to disagreements on how to define, categorize, and measure the mathematical knowledge for teaching (Hill, 2007; Kane, 2007; Schilling & Hill, 2007). More research addressing theoretical and practical issues related to the conceptualization and use of MKT were needed. These studies would assist in the development of measures to help validate claims that this type of knowledge is related to student performance in mathematics (Hill, Rowan, & Ball, 2005).

Mathematical Knowledge for Teaching

Mathematical knowledge for teaching consists of knowledge domains that are specific to the job of teaching mathematics to learners. Shulman (1986) was among the first to consider these various knowledge domains and formalize them in writing. In a concerted effort to explore these domains in mathematics education, Hill, Ball et al. (2008) broadened the scope of Shulman's (1986) pedagogical content knowledge (PCK) and content knowledge (CK) domains, rephrasing the latter as subject matter knowledge. The domain of subject matter knowledge consisted of categories such as common content knowledge (CCK), specialized content knowledge (SCK), and horizon content knowledge (HCK). CCK consists of the knowledge of the content that a typical adult is expected to have. SCK consists of the content knowledge that is common across fields or occupations. HCK consists of the knowledge of related concepts within the content area, but beyond the scope of the content being taught. PCK consists of domains in knowledge of content and students (KCS), knowledge of content and teaching (KCT), and knowledge of content and curriculum (KCC). KCS is about understanding why specific content is difficult for students. It also includes an understanding of the best ways to teach mathematical content. KCC suggests teachers should understand sequences in which content should be taught. By defining these domains of knowledge Hill, Ball, et al. (2008), provided a framework for conceptualizing the structure of MKT.

Theoretical considerations of each subdomain of mathematical knowledge for teaching have been discussed at length since Shulman's seminal work. Early conceptualizations of the function of subject matter knowledge suggest that this knowledge is important to teaching any subject because it contains the facts and concepts, as well as how those facts and concepts are organized (Shulman, 1986; Grossman, Wilson, & Shulman, 1989; Hill, Ball et al., 2008). This idea, though not novel (see Dewey 1910/1997), provides a starting point for identifying the categories that make up subject matter knowledge. In her 1990 analysis of teacher work samples, Ball delineates two categories of content knowledge in mathematics, knowledge of mathematics and knowledge about mathematics. Knowledge of mathematics consists of an understanding of facts and concepts whereas knowledge about mathematics consists of understanding why those concepts exist and how they should be applied. In an attempt to operationalize the concepts *knowledge of* and *knowledge about*, Hill et al. (2004) redefined these as *common content knowledge* and *specialized content knowledge*. Common content knowledge remained conceptualized as the average mathematical knowledge expectedly held by an adult. However, SCK became more refined by encompassing more strict definitions such as “building or examining alternative representations, providing explanations, and evaluating unconventional student methods” (p. 16). The other type of subject matter knowledge is *horizon content* knowledge. This type of knowledge includes an understanding how courses related to a particular subject matter are interconnected sequentially and concurrently (Shulman, 1986; Kreber & Cranton, 2000).

The breadth of research on pedagogical content knowledge provides indication of its importance not only in mathematics, but other fields as well (Ball, et al., 2008). However, like other domains of mathematical knowledge for teaching, an empirical definition of the PCK has been elusive. Ball and Bass (2000) broadly consider this PCK as knowledge for practice. Similarly, Niess (2005) provides a relatively broad but common definition that suggests PCK is the place where content knowledge and pedagogical knowledge meet.

The most current theoretical conceptualization of PCK is outlined in Ball et al. (2008). Here researchers delineate the subdomains of PCK, suggesting that KCS, KCT, and KCC can each be defined as separate constructs. KCS is an understanding of how students process certain content. KCT is an understanding of how to teach certain content. Though vague in their interpretation of knowledge of content and curriculum (Ball et al., 2008) this domain seems suggestive of having an understanding of how similar content is taught across different curriculum (Sleep, 2009). An illustration of the domains of MKT is in Figure 1.

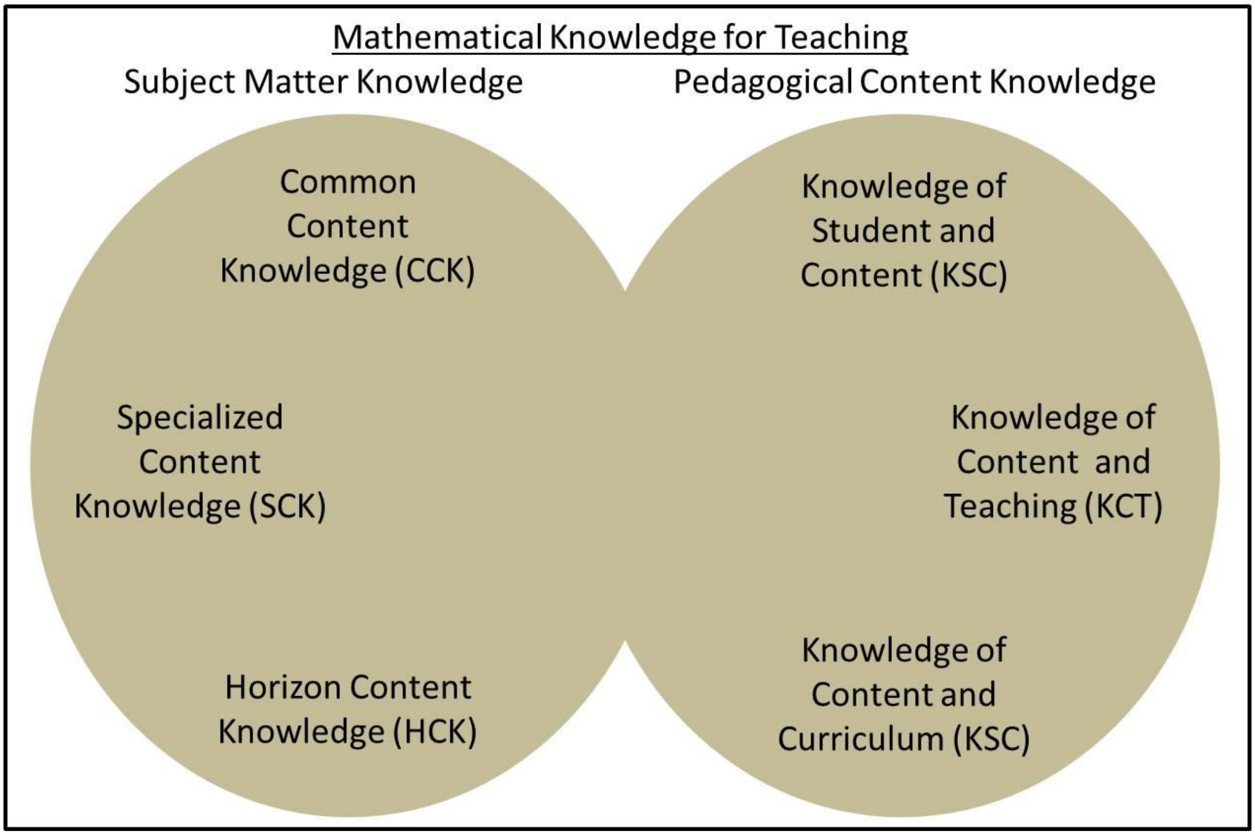

Mathematical Knowledge for Teaching
Subject Matter Knowledge
Pedagogical Content Knowledge
Common Content Knowledge (CCK)
Specialized Content Knowledge (SCK)
Horizon Content Knowledge (HCK)
Knowledge of Student and Content (KSC)
Knowledge of Content and Teaching (KCT)
Knowledge of Content and Curriculum (KSC)

Figure 1.Illustration of MKT domains adapted from (Ball, Thames, & Phelps, 2009).

Problems with the Construct.

Though these knowledge domains provide a framework for conceptually and empirically exploring mathematical knowledge for teaching, there still remains some ambiguity. Research to date has been unable to empirically delineate these domains (Hill et al., 2004). There are logistical challenges in determining whether these knowledge domains, as currently conceptualized, are the most appropriate for describing the knowledge that math teachers use. These challenges have yet to be overcome (Ball et al., 2008).

One of the major challenges in defining the MKT construct is the issue of construct representation. As previously mentioned the construct includes multiple subdomains. The relationships amongst these subdomains have not been thoroughly explored. For example, there exists no empirical evidence that demonstrates the relationship between specialized content knowledge and knowledge of content and teaching. Determining how these constructs relate to one another is an important aspect of ensuring that the construct is well represented. Although theory regarding MKT has evolved into these six subdomains, future studies may suggest that there may be more or less.

Another issue in conceptualizing the construct stems from its practical use. In developing this construct, Ball and colleagues (2008) researched the tasks of teaching mathematics. Each of the tasks represented teachers' use of MKT. However, it is unclear from the current research, whether the subdomains of MKT could be used differentially by teachers to complete these tasks. For instance, a teacher with more experience in the classroom may use his or her knowledge of student and content to assist a student with a problem that the teacher has seen in students of similar ability. On the other hand, a fairly new teacher could rely on specialized content knowledge to assist the same student in solving the problem. The effectiveness of either approach is determined by the student's future success on similar problems. If one domain can be compensated by another domain based in the same classroom situation, then is it really worth delineating the domains? Should we instead think more holistically about these knowledge domains? These questions can be answered by empirically examining teachers' use of MKT (Ball, et al. 2008).

Indirect Assessment of MKT

In-service Teachers

Teachers must have sufficient knowledge of the subject matter they teach. This level of sufficiency reaches beyond what would be required of the average adult (Borko et al., 1992). They are expected to have mastery of the content they intend to teach. Most would agree that without this knowledge, a teacher's effectiveness in educating k-12 students could be adversely affected (NMAP, 2008).Though it is widely understood that this "profound" knowledge of the mathematics is necessary for effective practice, researchers and teacher educators are still in search of what this knowledge consists of. Without a clear definition of MKT, making valid inferences about its effects on student achievement is not possible.

Though there has been a lack of substantial empirical evidence for making the connection between mathematical knowledge for teaching and student achievement, there has not been a shortage of anecdotal evidence concerning their relationship. Using interview questions developed by Ball (1988), Ma (1999) presented momentous qualitative research comparing U.S and Chinese elementary teachers on their understanding of mathematical concepts. The findings of the study suggest that the U.S. teachers possessed math skill that allows them to derive correct answers, but lack the "profound understanding" that is typified in the idea of mathematical knowledge for teaching. Identifying measures that are suggestive of this profound understanding helped to begin developing the construct of MKT.

Following the aforementioned descriptive research on MKT, researchers began further development of the MKT construct by using teacher production inputs as proxies for this knowledge (Hill, Blunk et al., 2008). Teacher production inputs such as mathematics courses attained, certification, and licensure status were considered indication that the teacher possessed MKT. These were linked to K-12 student achievement in mathematics (Begle, 1972, 1979; Hanushek, 1981, 1996; Harbison & Hanushek, 1992; Hill et al., 2005; Monk, 1993; Mullens, Murnane, & Willett, 1996; Rowan, Chiang, & Miller, 1997).

In a longitudinal study conducted by Rowan et al. (1997), classes of variables used as proxies for content knowledge in mathematics were found to have a small effect on student achievement. A meta-analysis conducted by NMAP (2008) showed further attempts to link student achievement in mathematics to the productivity inputs. However, productivity inputs such as teacher certification only provided indication that the teachers are certified according to some set of standards. Likewise, the course attainment variable only provided indication that the teachers passed a mathematics related course. Passage of a course could not indicate the degree to which knowledge was obtained as a result of the course. Neither certification status nor the number or type of mathematics courses taken by a teacher provided any indication of the specific knowledge that leads to K-12 student achievement in mathematics (NMAP, 2008).

Another proxy for measuring mathematical knowledge for teaching was passage of a licensure exam (Hill et al., 2005). This examination serves as an indication of attainment of requisite knowledge for teaching. It also serves as an accountability tool to ensure schools are hiring competent teachers. One such examination is the ETS-developed PRAXIS examination. This examination satisfies certification requirements in over 40 states. Some states also use their own certification exams which may be required en lieu of the PRAXIS. The emphasis that these examinations have on the mathematical knowledge required of teachers can vary. The PRAXIS mathematics portion focuses on content knowledge similar to SAT or NAEP measures. However, the California Basic Educational Skills Test and the Florida Teacher Certification Exam focus more on basic skills. These licensure and certification exams do not require the examinee to unpack mathematical ideas (Hill & Ball, 2004). That is, they do not require the examinee to discern mathematical procedures or to evaluate mathematical claims. These are two tasks that typify profound knowledge of mathematics for teaching.

These exams also lack congruence with the learning objectives of the programs from which these teachers get their training. The learning objectives of a teacher education programs are likely to consist of mathematical knowledge and skills that are more complex than the content knowledge assessed by the licensure exam. Research suggests that these content focused tests are not effective in measuring cognitively complex constructs such as mathematical knowledge (Nichols & Sugrue, 1999). The scores on these exams are likely to be over interpreted if a teacher education program uses the scores to address the domain of MKT. This is due to a lack of fidelity between the mathematical knowledge required for passing these exams and the MKT that teacher educators suggest preservice teachers should attain.

Once teachers are licensed or certified, mathematical knowledge for teaching is examined intermittently through professional development programs. These programs have become increasingly popular since the adoption of NCTM standards (Loucks-Horsely, Love, Stiles, Mundry, & Hewson, 2003; NMAP, 2008). Professional development has moved from focusing on the more generic instructional strategies to considering the content and pedagogical knowledge of teachers. However, rarely do school districts' allocate more than 10% of their budget for professional development. This figure includes all content areas, not just mathematics (Loucks-Horsely, et al., 2003). This illustrates that the organizations that create the standards or decide what teachers should know are not the same organizations that hold the purse strings for professional development (Ingvarson, 1998). Due to a lack of funding, professional development in mathematics education is often fragmented, episodic, and rarely meets the depth of knowledge suggested in current MKT literature (Ball, Luienski, & Mewborn, 2001). Therefore, once the breadth of knowledge for teaching mathematics is met for licensure sustained professional development of MKT ceases. This prevents researchers from developing measures of MKT that are more precise than the proxies that are typically used.

Preservice Teachers

There is a general expectation that preservice teachers should acquire knowledge for teaching mathematics during their preparation program. Programs in elementary education are expected to have curriculum designated for the development of MKT. However, a program's ability to determine the extent to which this knowledge is obtained is important to understanding the program's effectiveness on developing MKT.

Teacher education programs provide curriculum to assist preservice teachers in developing both content and pedagogical knowledge in mathematics. Students are tested on this material in various courses where curricula are based on standards of the program's accrediting body (e.g. NCATE). Though the standards are set by this accrediting body, the instruction methods are not specified. Therefore, the learning experiences used to meet these standards can differ across programs. This variation in learning experiences across programs presents an important challenge in determining the impact of teacher education on preservice teacher knowledge. Due to these differences across programs, each individual program must develop appropriate methods to assess the effectiveness of their programming.

Preservice programs typically adopt methods to assess all or a sample of their students in the knowledge or skill areas the program finds important. These areas are called learning outcomes, or objectives. Assessment of these learning objectives can occur anywhere within the curriculum. Programs that are interested in the skills or knowledge their graduates will carry into the workforce typically assess students near end of the students' academic experience. Results from these assessments can be used to determine how well students within the program are performing on the learning objectives (Erwin, 1991). These results provide indication as to what programmatic changes need to occur to improve student learning in those areas (Suskie, 2009).

The development of these program-level assessment plans has been sparked by national accountability movements. Historically, accreditation agencies have been only concerned with graduation rates, course attainment, and licensure exam pass rates. Similar to their work with in-service teachers, these proxies for the attainment of knowledge for teaching do little to inform the effectiveness of the teacher educational program. Recently, accountability efforts have increased the emphasis on rigorous assessment for teachers' preparation programs at many universities (Center for American Progress, 2010). Program accreditors are ensuring that students are being assessed. They are also interested in how institutions are using assessment findings to improve programs. To appease their interests, the assessment of MKT requires measures that are more precise than graduation rates, course attainment, and licensure exam pass rates.

Direct Assessment of MKT

Direct measures of MKT are those that examine the extent to which teachers can demonstrate this knowledge. Generally, these measures exist on a continuum that extends from objective assessment to performance assessment (Johnson et al., 2009). Objective assessments are those that restrict the examinee to specific number of response options that are provided by the test developer (e.g., multiple-choice). The focus of the assessment is on the examinee's ability to select the correct response to an item. Performance assessment requires the examinee to demonstrate the knowledge or skill being assessed through a performance (e.g., project or simulation). Both of these types of assessments have been used to assess mathematical knowledge for teaching. They each have pros and cons that affect their use in the assessment of MKT at the preservice program level.

Performance Assessment

The use of teacher portfolios is a commonly used form of performance assessment for preservice teachers. These portfolios include products that are representative of student performance on specific learning objectives. The portfolios not only provide an opportunity to assess preservice teachers on their ability to demonstrate their knowledge, but they also provide opportunity for reflection (Zeichner & Liston, 1996). Portfolios are used within individual courses, admission processes for student teaching, and throughout the duration of programs (Zeichner, 2000; Borko, Michalc, Timmons, & Siddle, 1997). They are also sanctioned by the National Board for Professional Teaching Standards as a procedure for obtaining certification.

There is a high degree of variability in the way portfolios are constructed and used (Zeichner & Wray, 2001). Borko et al. (1997) noted the lack of research that demonstrates systematic processes for creating and or implementing portfolios in teacher education. Consequently, there can be great variability in the effectiveness in meeting the purposes for which the portfolio is used (Wade &Yarbrough, 1996).

There are multiple types of portfolios that are used to meet different assessment needs. The learning portfolio is typically used to enhance reflection and provide information about growth. The credential portfolio is used to determine competency or attainment of standards set by the state or National Board (Snyder, Lippincott, & Bower, 1998). A showcase portfolio is used for the presentation of a preservice teacher's best work for purposes of employment. A type of showcase portfolio that has been used in preservice teacher assessment is the Teacher Work Sample. This portfolio consists of the teacher's best evidence that their k-12 students are learning. The products placed in the portfolio include rated observations of the teacher's classroom performance, rated observations of the teacher's ability to determine how well his or her students meet the learning objectives, and evidence of the teacher's ability to create assessments of student learning (McConney, Schalock, & Schalock, 1998).

Pratt (2002) conducted a study in which standards for teaching mathematics (NCTM, 1989) were aligned to the products provided in a TWS portfolio. Using a sample of 50 portfolios from k-12 preservice teachers at Western Oregon University the researcher addressed the alignment of the products with the standards. A weak alignment between the products within the TWS and the NCTM standards was found. This indicated that the portfolio did not include sufficient evidence that the teachers were meeting the standards. This lack of alignment illustrated a misconception about portfolio assessment when used in this context. The authenticity of the task or performance (e.g., classroom performance) does not guarantee fidelity between the product received (e.g. TWS) and the knowledge domain being assessed (Cizek, 1991). In this study the researcher attempted to use the work samples to satisfy a purpose for which they were not created.

Another issue impacting the variation of portfolios is determining what goes in them and who decides what should go in them (Johnson, et al., 2009). The *what* part can include observations, tests, lesson plans, and reflective statements. The *who* can be the preservice teacher or the program. Some portfolio systems may allow for the preservice teacher and the teacher educator to decide what goes in the portfolio. These decisions will affect the quality of the data that the program receives.

These variations in how portfolios are used make it difficult to ascertain their effectiveness in meeting the outcomes they claim to meet. For example, most research on the use of portfolios state that they increase student reflection. However, the researchers rarely examine the depth of that reflection or compare how the modification in the various aspects of the portfolio process can increase or decrease the achievement of reflective thinking (Borko et al., 1997; Wade & Yarbrough, 1996; Zeichner & Wray, 2001). This is an issue regarding the quality of the portfolio process. Without clear indication of the requisite quality of teaching portfolios, the effectiveness of their use in program-level assessment will remain uncertain.

Selected-Response Assessments

Over the past decade much of the research on mathematical knowledge for teaching has been conducted by researchers associated with the Learning Mathematics for Teaching Project. This project focused on developing measures of direct measures of MKT. Hill, et al. (2004) developed items to quantitatively measure mathematical knowledge for teaching with in-service teachers. These items differ from the items one would typically see on a licensure examination in that they attempt to assess several domains of MKT, instead of focusing on general subject matter knowledge.

The Content Knowledge for Teaching Mathematics (CKT-M) items emerged out of a project that focused on developing test items for measuring the knowledge required for teaching mathematics (Hill et al., 2004). These items were created for use with in-service teachers who have completed professional development programs for improving knowledge and skill for teaching mathematics. The items purportedly assess mathematical knowledge for teaching as conceptualized by the test developers. The theoretical underpinnings for creating items that assess mathematical knowledge for teaching stem from the work produced by Ball and Bass (2000), Grossman (1990), and Shulman (1987).

The developers created the items to focus on three content areas: 1) numbers and operations, 2) geometry, and 3) patterns, functions, and algebra. They also explored two types of knowledge domains used by teachers. 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 mathematics knowledge expected to be held by an average adult. Specialized content knowledge consists of an understanding of mathematical concepts unique to 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 these mistakes are made, as well as identifying students' problem solving strategies.

Hill et al. (2004) developed and piloted three forms of selected response items written to represent the aforementioned knowledge domains and content areas, excluding geometry items. These items were piloted on samples of in-service teachers participating in California's Mathematics Professional Development Institute. An exploratory factor analysis (EFA) using principal axis factoring with promax rotation was conducted on each form to identify the factor structure underlying the measure. The geometry items were removed to reduce the complexity of the model. Results indicated that a two-factor model best fit Form A, while a three factor model best fit Form B and C. Although the factors did not align as theorized due to the multidimensionality of items, the developers concluded that a three factor model was adequate for explaining the data. An examination of the factor loadings revealed a relationship among items within the areas of 1) knowledge of content in numbers and operations, 2) knowledge of student and content in numbers and operations, and 3) knowledge of content in patterns, functions, and algebra.

Using the same data, Hill et al. (2004) also conducted a bi-factor analysis to explain the relationship between general and specific factors related to mathematical knowledge for teaching. They found that a substantial number of items loaded onto a general math knowledge factor. More specifically, 67-77% of the variance in responses to items across the three forms was explained by this general factor. Items also loaded differentially onto the specific factors representing items written in a combination of content and knowledge domains. Multidimensionality was found, with no firm patterns of loadings across the three forms. 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.

Using the same forms from the bi-factor analysis, Hill et al. (2004) further investigated the properties of each item using IRT methods. Adequate reliabilities were found for numbers and operations items in knowledge of content and knowledge of student and content domains. Adequate reliabilities were also observed for items measuring patterns, functions, and algebra in the knowledge of content domain. However, multidimensionality amongst several of the items was apparent. Again, this finding underscores the need for more psychometric work with the instrument.

The CKT-M items have been used to measure MKT growth of in-service teachers participating in professional development programs. Results from California's Mathematics Professional Institutes indicate that the items could capture growth and that the growth is related to program duration (Hill & Ball, 2004). It has also been used to link student achievement to teachers' knowledge for teaching mathematics (Hill, et al., 2005). Their study mixed quantitative and qualitative research methods to address the relationship between teachers MKT and K-12 student achievement. The results indicated that growth in MTK as measured by the CKT-M was a significant predictor of first and third grade gains in students' mathematics achievement.

Though these studies have provided insight into the items' usefulness with in-service teachers, there have been few inquiries into their psychometric properties with preservice teachers. It is good psychometric practice to conduct necessary steps to insure that CKT-M items are suitable for the population in which inferences are to be drawn from (APA, 1999). Therefore, evidence needs to be gathered for preservice teachers. Recent studies relating mathematical beliefs and quality of instruction to scores on the CKT-M items did not attempt to gather evidence (Swars et al. 2007; Sleep, 2009).

In a longitudinal study, Swars et al. (2007) examined the relationship between preservice teachers' beliefs about mathematics and mathematical knowledge for teaching. Their beliefs about teaching and learning mathematics were assessed through a survey. MKT was assessed using the items developed by Hill et al. (2004). Results indicated a significant correlation between growth in scores on the CKT-M items and mathematical beliefs. Though Swars et al. (2007) were able to provide some insight into how MKT can be used to discuss mathematical belief systems, there were several methodological flaws that limit the veracity of the results. Dimensionality of the CKT-M items was not addressed. Previous research has indicated the existence of multidimensionality in the CKT-M items (Hill et al., 2004). When items measure multiple dimensions the interpretations of scores on those items should be adjusted accordingly (Reckase, 1979). Swars et al. (2007) did not address how the reliability of the items could be affected by this multidimensionality. The authors only reported the reliability of the items as they were obtained during development by Hill and her colleagues. However, reliability is not a property of a test, and must be addressed during each administration (Thompson & Vacha-Haase, 2000).

Sleep (2009) examined the relationship between scores on the CKT-M items and preservice teachers' ability to design and steer instruction in the classroom. This study used the CKT-M items to rank order their participants for selection into a follow up group of interviews. Similar to Swars et al. (2007), reliability and dimensionality were not addressed for their sample of participants. When reliability estimates are biased, the rank ordering of participants based on their scores may not be appropriate. Unreliable scores are likely to produce unreliable rank orders (McDonald, 1999).

Russell et al. (2010) examined whether the CKT-M items were appropriate for a preservice teacher sample. Data from 988 preservice teachers were analyzed using confirmatory factor analyses (CFA). This analysis allowed the researcher to test the fit models for scoring the data. Results indicate that an appropriate factor structure could not be found for scoring the data. The multidimensionality of the data, as well as the item dependencies caused by the testlet structure was suggested to be the cause of poor model fit. Similar results were found in a study of the reliability of these items with a preservice sample by Gleason (2010). Citing a lack of independence of the items, the author reasoned that the reliability of the items with preservice teachers is significantly lower than what would be appropriate for interpretation.

The issue of multidimensionality does not bode well for the needs of program assessment. The purpose of program assessment is to identify areas where students are doing well or not and to make changes accordingly. If we cannot delineate the construct of MKT by creating items that can provide some reliable indication of proficiency in specific areas of MKT, then the purpose of program assessment and improvement will not be met. The program will be unable to make informed changes that will improve the attainment of knowledge that is not specified.

The lack of tools for the purpose of assessing preservice MKT for program improvement presents a challenge to teacher education programs. Without proper tools for assessing their students, making well informed decisions regarding changes to teacher preparation programs becomes difficult. However, in this challenge lies an opportunity to develop measures specifically for assessing pre-service teachers for the purpose of program assessment and improvement.

Validating Measures of MKT for Preservice Teachers

Validation is the process by which the inferences made from scores on a test become trustworthy (Messick, 1989). When measuring a construct, the researcher needs to be confident in his or her interpretation of the test score. This confidence is gained by obtaining validity evidence. It typically requires several research efforts to obtain adequate validation evidence. This is because validity is a matter of degree, not a dichotomous decision (Benson, 1998).

In the case of CKT-M measures, validation evidence is still being sought for in-service teachers (Hill, Dean, & Goffney, 2007). This lack of construct validity continues to be a challenge to developing measures of MKT for use in preservice program improvement. Benson (1998) offers a framework for addressing construct related validity issues. There are three stages within this framework: substantive stage, structural stage, and external stage. The substantive stage involves theoretically and operationally defining the construct. If the construct is well defined, items can then be written to clearly represent the construct. The structural stage involves examining the interrelations amongst the items within the measure. This is done to determine the extent to which items relate in theoretically meaningful ways. The external stage involves examining the relationship between the measure and theoretically relevant variables.

Each of the aforementioned stages allows researchers to build evidence for the inferences they wish to make from their measures. Studies associated with the development of thick-M items addressed some of the stages within this construct validation framework. For example, developers of the CKT-M items built a theoretical framework of in-service teacher MKT that was operationally defined by their items (Hill et al. 2004). They conducted factor analyses to examine the interrelationships of the items. Later, they explored the relationship between their measure of MKT and a measure of mathematical quality of instruction (Hill et al., 2008). Jointly, these research efforts provide some validation evidence for the inferences made regarding in-service teachers' mathematical knowledge for teaching.

In studies with preservice teachers, the validity evidence for use of thick-M items to assess MKT was not apparent. At a surface level, the items appeared to measure the MKT construct in such a way that they could be used for preservice teacher assessment. At the time of the study conducted by Russell et al. (2009), there existed no literature that suggests theoretically or empirically that the items would not perform similarly for in-service and preservice populations. However, psychometric investigations into the factor structure of the items suggested that the knowledge domains underlying responses to the items were dissimilar to that of in-service teachers (Russell et al., 2010). Furthermore, the reliability of the items when used in a preservice population was less than adequate (Gleason, 2010). Lastly, the items did not function as expected with preservice teachers at different levels of their program. The items were not sensitive enough to capture the effect of the mathematics education curriculum on preservice teacher MKT (Russell et al., 2011). This indicated a lack of validity evidence at the external stage of construct validation. Overall, the validity evidence obtained in these studies suggested that the items should not be used for preservice teacher program assessment.

Though the CKT-M items did not perform well for the purpose of preservice program assessment, the process of validation remains a viable framework for the development of preservice measures of MKT. Beginning with the construct, developers of new items can focus on defining aspects of the construct that are most relevant to a preservice teacher program. Items can be developed to assess the aspects of MKT that should be gained as a result of the program's mathematics education curriculum. By focusing only on those aspects of MKT, the item developers can further clarify the domain of interest as illustrated in the substantive stage of construct validation.

By clearly defining the substantive domain of the construct under investigation, the preservice program can then create items that directly assess the construct under investigation. The factor structure of those items can then be tested to determine how well they fit the model under which they were created. The fit of the model could provide evidence that the items are measuring what they were developed to measure. This is the essence of validity. However, further validation evidence could be sought by identifying whether the items perform as expected with different populations (e.g., beginning preservice teachers versus graduating preservice teachers). Such evidence would provide information about the external stage of Benson's framework.

Benson's (1998) framework is thorough in its approach to addressing construct validity. However, it does not explicitly take into consideration qualitative forms of inquiry. Qualitative methods can be used to explore the same constructs measured by quantitative methods. However, the process of addressing the validity of inferences made from that exploration differs.

Generally speaking, qualitative studies conceptualize validity in terms of *credibility, transferability, dependability, and confirmabilty* (Merriam, 2009). Each of these terms relate to the process of making the data in such a way that it best represents the reality of the experience, environment, and participant under investigation (Richards, 2005). Though the phrase "making the data" is foreign to users of quantitative methods, it is an important concept in qualitative methods. This phrase suggests that the researcher is the instrument by which data is obtained. Investigator bias is a major threat to this process of "making data. "Users of this method of inquiry must attempt to reduce this threat to their credibility just as users of quantitative methods must reduce the impact of experimenter bias on validity. Qualitative researchers achieve this by attempting to remove personal bias so that the data reflect the reality of their object of inquiry.

Typical ways of obtaining internal validity evidence include *triangulation* and *member checking* (Merriam, 2009). Triangulation involves the use of multiple methods of data collection, multiple sources of data, and multiple investigators. Member checking occurs when investigators obtain feedback from respondents regarding their interpretation of the data. Both of these approaches allow researchers to address their data's proximity to reality.

External validity addresses the extent to which results are generalizable in quantitative studies; however, in qualitative research the term *transferability* is used to describe how applicable the results are to other similar situations (Merriam, 2009). Consequently, qualitative researchers must ensure that they do not go too far in controlling the variables in their study. Doing so would create a less authentic situation that may not be transferable.

Two ways of obtaining external validity evidence using qualitative methods include *thick description* and *maximum variation* (Merriam, 2009). Thick description is a process of describing the context of the study in great detail. This allows other researchers to explore the depth to which the findings transfer. Maximum variation is a process by which participants are purposefully selected to maximize the differences amongst the participants in a sample.

Although much of the process of obtaining validity evidence for qualitative data differs from Benson's framework, there is one aspect that is similar. Merriam (2009), like Benson (1998), suggest using multiple methods to address validity. They further note that the overlap of the methods can provide some convergence of findings. This convergence would strengthen validity of the inferences obtained from the data.

Most qualitative studies examining MKT have focused on illuminating the experiences of in-service teachers in order to explain their use of MKT (Cohen, 1990; Heaton, 1992; Putnam et al., 1992). Through the use of observations, in depth interviews, and open ended questions researchers have been able to describe deficits in mathematics teachers' ability to provide effective mathematics instruction. Heaton (1992) used observation techniques to examine the use of mathematics content knowledge. A case study situated in a classroom revealed that the in-service teacher was able to make the connection between the presentation of the subject matter and the goals of the lesson. However, the teacher's lack of subject matter knowledge reduced her effectiveness in ensuring that the students were meeting the lesson's learning outcome. Putnam et al. (1992) found similar results using the case study method. Though these studies provide a description of MKT use that is transferable, they do little to help improve our ability to measure the construct.

Cognitive interviews have been used in studies to assist in the instrument development process. This process typically involves respondents speaking aloud while they respond to items on an instrument (Strack & Martin, 1987; Tourangeau, 1987; Tourangeau & Raskinski, 1988). They may also be asked to respond to questions about the items. This process is conducted to determine the relevance and the clarity of the items on an instrument. Identification of the relevance and clarity of items through the use of cognitive interviews can contribute to the provision of construct validity evidence for an instrument (DeVellis, 2003).

Research has pointed to several benefits in using cognitive interviews with respondents. They allow the researcher to gain valuable insights into the thought processes used by respondents in answering the items. This includes information about how the respondent interpreted the item and provided rationale for their response (Collins, 2003, Drennan, 2003; Williamson & Raynard, 2000). These insights have assisted in identifying items with poor wording or items that do not reflect the knowledge or skills they were intended to illicit. Identifying problematic items is an important aspect of instrument development (Beck &Gable, 2001; DeVellis, 2003).

Though this process of obtaining content validity bodes well for the instrument development process, few guidelines exist in determining how data from cognitive interviews should be analyzed, interpreted, and used (Knafl et al., 2007). Drennan's (2003) review of cognitive interview literature noted the subjectivity used in the process of analyzing this type of data. For example, Willis, Royston & Bercini (1991) noted issues with their items that were related to ambiguity, ordering and relevance. However, their documentation did not note how they analyzed the information gathered through the cognitive interview process nor how they determined how best to use their results to make changes to items.

Counter to the general discussion in previous research on how cognitive interview data is analyzed, Knafl et al. (2007) provides a framework for systematically analyzing this data. The researchers used analytic coding techniques that are specific to the process of conducting qualitative research (Richards, 2005). They also developed a detailed process of categorizing problems with their items, as suggested by literature (Willis et al., 1991).

Using cognitive interviews to gather convergent validity evidence in instrument development research can be discussed in the context of a research methodology known as mixed-methods. The tenets of mixed methodology suggest that quantitative and qualitative methods can be combined to answer research questions (Creswell & Plano Clark, 2007). By combining the methods, the researcher is able to strengthen inferences that may otherwise be weak. In the case of the quantitative focused analysis, the voice of the participant typically goes unheard, thus neglecting important validation evidence (Creswell & Plano Clark, 2007). On the other hand, qualitative focused research typically provides the inquirer with access to the participants' voices. However, generalizing beyond those voices becomes a logistical and theoretical problem (Merriam, 2009). Mixing the two methods provides a viable solution to reducing the impact of either method's limitations. This approach has been used effectively in theory building and instrument validation (Aldridge, Fraser, & Huang, 1999; Myers & Oetzel, 2003; Hill, Blunk, et al., 2008). Through the use of this method Hill, Blunk, et al. (2008) were able to identify a relationship between MKT as scored by the CKT-M items and mathematical quality of instruction.

Mixed-methods approaches have not gone without their skeptics who suggest that the methods therein are incompatible due to their reliance on separate sets of assumptions (Guba, 1987; Johnson & Onwuegbuzie, 2004). One major point of contention is the quantification of qualitative results for purposes of mixing the data. This process inherently violates assumptions of quantitative analysis, which require that data are obtained independent of the researcher's perspective. This violation is typically seen in instrument development research (Morgan, 1998). However, this violation can be avoided if care is taken to preserve the integrity of either method during mixed methods research (Hanson, Creswell, Clark, Petska, & Creswell, 2005). As suggested by Sale, Lohfeld, and Brazil (2002), this process begins with situating the phenomenon under investigation within the paradigm most appropriate for the research questions and then following the assumptions of each paradigm prior to interpreting the mixed results. It is this process by which the research herein will be conducted.

Summary

Teacher preparation programs have a duty to demonstrate effectiveness in preparing students to become competent and skillful mathematics teachers (NCATE/NCTM, 2003). This competence and skill can be demonstrated by students' use of the various domains of mathematical knowledge for teaching (Ambrose, 2004; Ball, Sleep, Boerst, & Bass, 2009; Kejander, 2007). Many programs are currently in search of instruments that can provide them with an indication of preservice mathematical knowledge for teaching (Russell et al., 2010). Several types of instruments are being used to fulfill programs' need to assess preservice teacher knowledge. However, in most cases, the instruments fail to provide adequate information that can be used to improve the program, and thereby improve student learning (Hill et al., 2004; Wade et al., 1996; Zeichner et al., 2001).

Unfortunately, the pace of research on the development of adequate instruments for assessing preservice teacher knowledge has not kept up with the advancement of theory concerning the domains of MKT. Much of the empirical research on instrument development has suffered from a lack of congruence among researchers on how best to define or measure the construct (Hill, 2007; Kane, 2007; Lawrenz & Toal, 2007). However, this lack of congruence provides an opportunity for preservice teacher programs to define and assess MKT using instruments that are tailored to their program's specifications. By defining MKT in accordance to the program's learning objectives, the program can then create measures to assess those learning objectives. The validity of the inferences made from the measure can then be addressed empirically. This research empirically addresses the validity of inferences from a measure of MKT by answering the following questions:

1. How do the PMKT items perform?
    - What factor structure is plausible for the data?
    - How does item difficulty and item discrimination vary?
    - How effective are the distractors?
    - What level of reliability is demonstrated by these new items when used with this pre-service teacher sample?
2. How do groups of pre-service teachers at different academic levels compare on their aggregate scores?

3. How do pre-service teachers conceptualize (i.e. think about) MKT and how does this conceptualization relate to the development of items for assessing MKT?
4. What level of face validity do the new items created to assess mathematical knowledge for teaching have with a pre-service teacher sample?

CHAPTER THREE

METHODS

Research Investigation

This mixed methods study blends data collection and analysis methods consistent with the philosophical underpinnings of explanatory mixed methods designs. In this design, the researcher uses both quantitative and qualitative methods to answer research questions regarding item functioning and construct validity in the instrument development process. Research questions 1 and 2 were answered with quantitative methods, whereas questions 3 and 4 were answered using qualitative analysis. A major assumption of this blending of methods was that qualitative data would provide useful context to the quantitative data obtained in the first stage (Creswell & Plano-Clark, 2006). This context was obtained by way of meticulously describing participants' experiences during the qualitative component of the study (Merriam, 2009).

Participants

Participants in this research consisted of 665 undergraduate students attending a Mid-Atlantic rural university. Participants were either enrolled in a preservice teachers program (n=396) or were members of a group of volunteers (n=269) from the University's subject pool. The student population at the University is 84% White, Non-Hispanic and 61% female. However, the preservice teachers program was 96% female, and 90% White, Non-Hispanic at the time of the study. Participants in this study closely matched the demographics of the preservice teacher program with 98% female and 93% White, non-Hispanic. Students in the preservice teachers program are required to complete math content courses. For the purposes of this study, data were collected from preservice teachers in courses MATH 107, MATH 207, and ELED 433. These were pivotal courses where mathematical knowledge for teaching was to be gained. MATH 107 and MATH 207 were core requirements for preservice teachers across all concentrations within the preservice program (e.g. Elementary Education or Early Childhood Education). ELED 433 was a required course only for undergraduate preservice teachers who were concurrently obtaining a Master of Arts in Elementary Education.

Math 107 focused on preservice teacher development in numbers and operations. It was the first required math course for students. Participants (n = 249) from this course provided an indication of preinstruction baseline levels of mathematical knowledge for teaching. These participants were assessed at the beginning of the course, prior to receiving any math instruction.

Math 207 focused on preservice teacher development of effective problem-solving strategies. These strategies were based on problems from content areas of mathematics such as data analysis and probability. Participants (n = 92) from this course provided an indication of intermediate levels of MKT having completed at least two courses designed to improve this knowledge. These participants were assessed at the end of this course.

ELED 433 emphasized preservice teacher development in children's mathematical learning and pre-numerical stages through the acquisition of advanced numerical processes and operations and connections to geometric and algebraic reasoning. This course represented the culmination of MKT for the undergraduate preservice teachers program. Participants (n = 55) at this level were expected to be advanced in their mathematical knowledge for teaching. The participants were assessed at the end of this course

While all students in each of the pivotal courses were required to participate in this study as part of the ongoing program assessment process, participants from the non-preservice teacher comparative group (n=269)were recruited through the University's subject pool. Students accepted from the subject pool were not enrolled in the preservice teacher education program and were included as a comparison group for the purpose of examining pre-existing differences on Mathematical Knowledge for Teaching between those who self-selected into the teacher education program and those who did not.

Instruments

Demographic Questionnaire. This questionnaire was used to gather participants identifying data such as their ID, age, academic level, current math course, and practicum experience. This data was used to categorize participant into groups that reflect their training in mathematics education. Participants received this demographic questionnaire prior to the administration of any cognitive assessment items. Demographic questions are presented in Appendix A.

Preservice teachers' Mathematical Knowledge for Teaching items (PMKT). The PMKT consists of 23 items developed by subject matter experts in the field of mathematics education. Each item was developed to assess specialized content knowledge domain of mathematical knowledge for teaching. This knowledge is defined by the student learning objectives for the teacher preparation program. Each item developed in conjunction with this study focused on the numbers and operations content area. Prior to analyzing the items, the aforementioned content experts participated in a content validation exercise to address the substantive stage of Benson's (1998) framework for construct validation. First, research on the operationalization of the MKT construct was consulted in developing the program's learning objectives (Ball et al., 2008). Each of the newly developed items was then mapped to the program's learning objectives. The mapping of the items to the MKT construct via learning objectives was conducted by two content experts who served as faculty members in the teacher education program. Consistent with content validity research, these experts were asked to independently rate each item as relevant or not relevant to each of the learning objectives (Crocker & Algina, 1986). Next, the experts were required to obtain consensus for items on which there was disagreement. There were some objectives to which no item could be mapped. This process of mapping the items was overseen by a program assessment specialist. The learning objectives and corresponding items are listed in Table 1.

Table 1. Content Map of PMKT items

Learning Objectives	# of Items	Item # on PMKT Assessment
Objective 1: Evaluating a K-8 student's mathematical work or arguments to determine if the ideas presented are valid	20 (87% of Test)	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22
Objective 2: Developing mathematically appropriate responses to students' "why" questions	0 (0% of Test)	--
Objective 3: Finding an example to make a specific mathematical point	3 (13% of Test)	3, 9, 18
Objective 4: Recognizing the mathematical ramifications of using a particular representation	7 (30% of Test)	4, 9, 10, 11, 12, 15, 16
Objective 5: Linking representations to underlying ideas and to other representations	14 (61% of Test)	3, 4, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 22
Objective 6: Connecting a mathematical topic to more- and less-advanced related topics	0 (0% of Test)	--
Objective 7: Choosing and developing useable definitions for mathematical concepts	0 (0% of Test)	--
Objective 8: Using mathematical notation and language and critiquing its use	7 (30% of Test)	1, 5, 6, 8, 12, 22, 23
Objective 9: Inspecting equivalencies by determining if two solutions that appear to be very different are actually equivalent.	9 (39% of Test)	1, 4, 6, 9, 10, 11, 15, 17, 18
Total Test	23 (100% of Test)	1-23

All items are presented in Appendix A. Each item was developed to simulate the process of using MKT in a classroom setting. The stem for each item provided a context of a classroom setting in which MKT is used. Participants were asked to select a response by using their specialized knowledge of mathematics. Distractors for each item were developed to assist in diagnosing common misconceptions in the use of MKT, thus providing a formative evaluation component within each item. Unlike previously used measures of MKT, the items developed in this study were developed to assess the MKT related learning objectives of the teacher preparation program. Therefore, the MKT construct was defined by the learning objectives of the teacher preparation program. This would allow the preservice program to use the items to determine the extent to which the program is assisting preservice teacher in obtaining the learning objectives. Also, by abandoning the testlet structure of items used in previous studies (Hill et al., 2004) test scoring and the assessment of psychometric quality were greatly simplified.

Student Opinion Survey (SOS). This 10-item self-report measure was used to examine student motivation in responding to the mathematical knowledge for teaching items (Sundre & Moore, 2002). Level of student motivation was identified by two factors measured by the scale, *effort* (5-items) and *importanc*e (5 items). The items were five-point Likert-type. The lowest possible score on an item is 1 and the highest possible score on an item is 5. Higher scores on items related to effort and importance were indicative of higher motivation. This scale was provided at the end of the assessment session. Consistent with previous studies, participants scoring less than 10 on the effort subscale were removed from the analysis (Wise et al., 2006). The survey questions are presented in Appendix B.

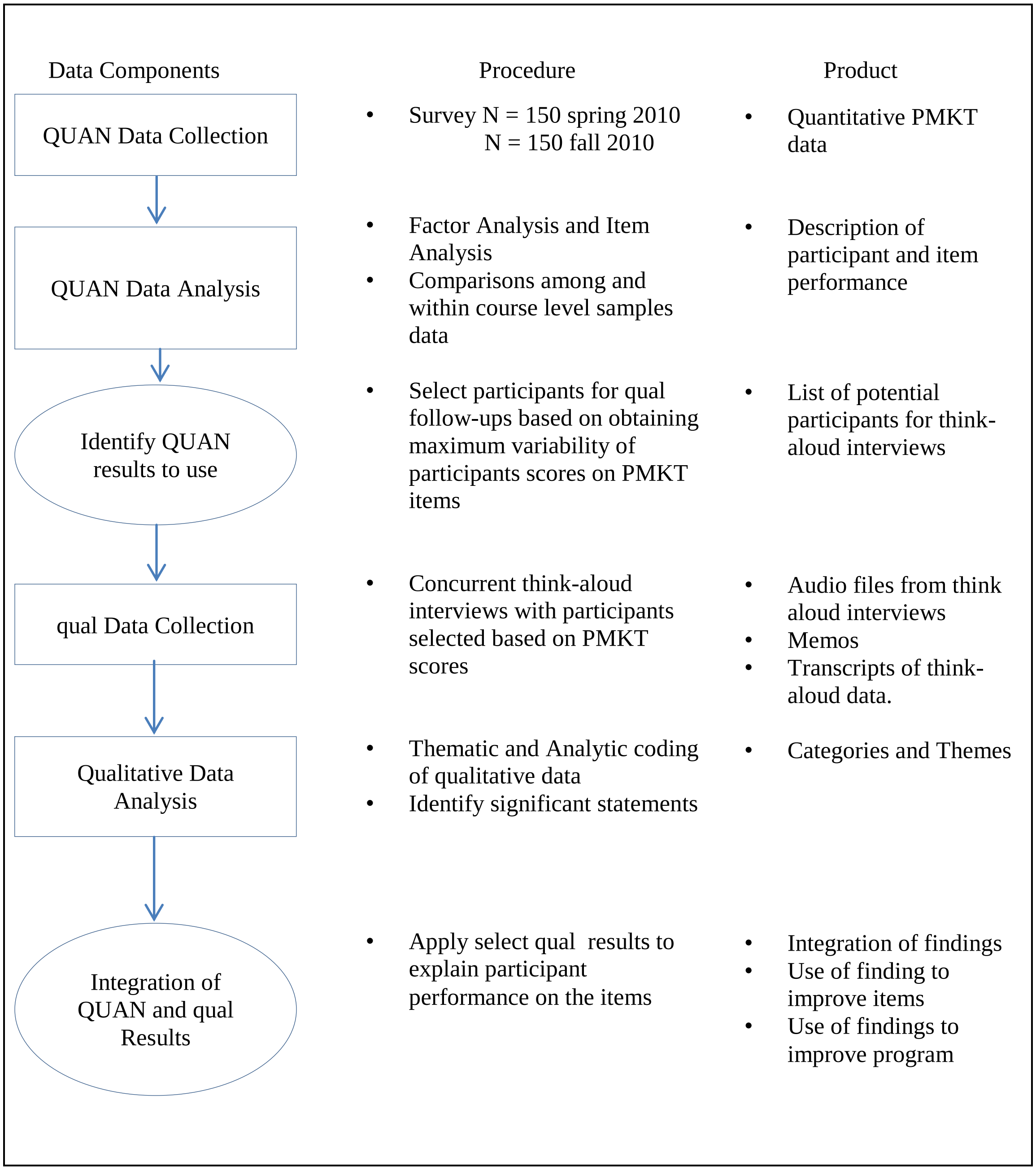
Think-Aloud Protocol. The think-aloud protocol was used to prompt the interviewee for verbal responses about their cognitive processes while completing test items (Ercikan et al., 2010). The protocol used in this study was based on the work of Sudman, Bradburn, and Schwartz (1996), which utilized retrospective debriefing through the use of open ended questions to assist the inquirer in obtaining clarification or more information. The think-aloud protocol is presented in Appendix C.

Research Design

A cross sectional design was employed to allow for data collection at pivotal points of the mathematics curriculum within the preservice teacher program. Preservice teachers enrolled in these pivotal levels of the program's math courses were mandated to participate in this study as part of ongoing program assessment. Each preservice teacher was administered a demographic survey, the PMKT items, and the SOS. To include students at all levels, this data collection spanned two semesters. Preservice teachers enrolled in Math 207 and ELED 433 courses for the Spring 2010 semester were administered these items at the end of the course. A baseline group of preservice teachers enrolled in Math 107 were assessed at the beginning of the fall 2010 semester. Also, during the fall 2010 semester a comparative group of non-teacher education students were selected from the psychology volunteer pool to provide a comparison to students who self-select into the preservice teacher program. All comparative group participants were administered the same instruments as the preservice teachers. The research design is illustrated in Figure 1.

The method of administration consisted of walk-in administration of the PMKT. The PMKT items were administered in a proctored computer lab. This was the assessment process adopted by the program to conduct program-level assessment. Program assessments have been administered this way since fall 2005. However, for this administration, the PMKT items were provided using a paper copy while the answers were recorded on the computer.

Once the quantitative data were obtained, it was then analyzed to examine the psychometric properties of the items and instrument. Next, results from the quantitative analysis were used to identify participants for the collection of qualitative data. Participants were selected in order to maximize the variability of participants' scores on the PMKT items within each course level. For example, participants with the highest and lowest scores on the PMKT items within each level of math course were sought for participation in the cognitive interview. This type of stratified purposeful sampling allows for the selection of participants based on a characteristics that can be easily used develop subgroups, such as high or low scorers (Patton, 1990). The identified participants were then solicited to participate in cognitive interviews. Eight volunteers per course level were solicited. Results from the qualitative analysis were used to provide contextual validation evidence for the functioning of the items.



Data Components	Procedure	Product
QUAN Data Collection	• Survey N = 150 spring 2010 N = 150 fall 2010	• Quantitative PMKT data
QUAN Data Analysis	• Factor Analysis and Item Analysis • Comparisons among and within course level samples data	• Description of participant and item performance
Identify QUAN results to use	• Select participants for qual follow-ups based on obtaining maximum variability of participants scores on PMKT items	• List of potential participants for think-aloud interviews
qual Data Collection	• Concurrent think-aloud interviews with participants selected based on PMKT scores	• Audio files from think aloud interviews • Memos • Transcripts of think-aloud data.
Qualitative Data Analysis	• Thematic and Analytic coding of qualitative data • Identify significant statements	• Categories and Themes
Integration of QUAN and qual Results	• Apply select qual results to explain participant performance on the items	• Integration of findings • Use of finding to improve items • Use of findings to improve program

Figure 2. Illustration of Explanatory Mixed-Method Design

Analytic Strategy

Research Question 1: How do the PMKT items perform? One of the first steps in evaluating item performance is to address their intercorrelations (DeVellis, 2003). This step coincided with the structural stage of Benson's (1998) framework, which suggested that factor analysis could be used to examine the factors or dimensions that underlie the interrelationships among the items. Mplus statistical software (Muthén & Muthén, Los Angeles CA) was used to test the dimensionality of the PMKT items under a confirmatory factor analysis framework. This process involved testing the hypothesis that the items were measuring only one factor. The fit of this unidimensional model was examined using χ^2, CFI, TLI, RMSEA, and WRMR fit indices. Poor fit of this unidimensional model was determined by the cutoffs for fit indices cited in Yu and Muthén (2002).

As a follow up to the poor fit of a unidimensional model, an exploratory factor analysis (EFA) framework was used. The number of factors retained from this EFA was determined by the Eigenvalue greater than 1 rule, scree plot, and parallel analysis. The eigenvalues indicated the amount of variance in a set of items explained by a factor. The scree plot illustrated the relationship between eigenvalues and the amount of variance explained by each factor. The parallel analysis involved a comparison of the observed eigenvalues to eigenvalues from a random data set, allowing one to determine an appropriate number of factors to extract (e.g., Fabrigar, Wegener, MacCallum, & Strahan, 1999). Best practices indicated that a combination of these procedures be used for determining the number of factors to retain (Bandalos & Finney, 2010). After extracting factors based on the aforementioned criteria, the factors were further analyzed through an examination of the factor loadings and communalities. Communalities provided an indication of the proportion of a variable's variance that could plausibly be explained by the underlying factors. Factor loadings illustrated the unique relationship between the variables and the factors after removing its relationship with other variables. Items with communalities less than .20 and factor pattern loadings less than .30 were considered for revision or removal.

Following the factor analysis, an item analysis was conducted to further explore the functioning of the items. By conducting the factor analysis first, the item statistics used to address item functioning could then be interpreted in the context of the data's factor structure (Reckase & McKinley, 1991). The data from the PMKT items were analyzed using SAS (SAS Inc., Cary NC). Analysis of these items provided further indication of the quality of each PMKT item. Item difficulties and discriminations were used to characterize the items. Characteristics of the items were analyzed for the entire sample, as well as for subgroups based on course level.

Item difficulty is determined by *p* values, which represent the proportion of examinees correctly responding to an item. This was calculated by dividing the number of correct response on an item by the total number of responses for the item. *P* values range from 0 to 1. Low values (less than .30) indicate higher difficulty, whereas high values (greater than .80) indicate lower difficulty. In this study, items with *p* values less than .30 and greater than .80 were considered for removal and revision.

Item discrimination is calculated using the point biserial correlation. This correlation reflects the relationship between the dichotomous 0 or 1 item score and the total score on the test. Point biserial values range from -1 to 1. Higher positive values represent a closer relationship between correctly responding to an item and higher scores on the test. Low values indicate that scores on the item are ambiguous indicators of examinees performance on the test. Negative values indicate that incorrectly responding to the item is related to higher scores on the test. In this study, items with point-biserial correlations less than .20 were considered for removal and revision.

A distractor analysis was also conducted to determine the quality of each item's response options. Distractors are incorrect response options. The distractors for the PMKT items were developed to address misconceptions amongst the participants. Distractors not chosen by the participants were not providing any information regarding participants' misconceptions. This is an indicator that the distractor was poorly written. Poorly written distractors can decrease the quality of the item and adversely impact the meaningfulness of test scores (Kaplan & Saccuzzo, 1997). Replacing non-distractors can improve scales by increasing item difficulty (Cizek & Day, 1994). Items for which no distractors were chosen were considered for removal and revision. For example, consider a multiple-choice item with four responses options. One of the response options is correct. The others are incorrect, but equally effective as distractors. For this item there is 25% probability of a correct response being obtained by guessing. However, if two of the distractors for this item were poorly written distractors, then the probability of correctly guessing increases. This type of item should be revised to improve the item's quality.

Research Question 2: How do groups of pre-service teachers at different academic levels compare on their aggregate scores? External validity evidence was obtained through group differentiation on the PMKT items. Group comparisons were made based on courses in which preservice teachers are enrolled. Preservice teachers enrolled in ELED433 were compared with preservice teachers in MATH107, MATH207, and non-preservice teacher participants. MATH207 were compared with MATH107, ELED433, and non-preservice teacher participants. Analysis of covariance was conducted to determine whether differences in PMKT total scores across levels of preservice math courses existed after controlling for SAT MATH. Significance tests and effect sizes were used to describe the differences in mathematical knowledge for teaching among groups of students. Preservice teachers in courses at higher academic levels were expected to perform better on items than preservice teachers in courses at lower academic levels.

Research Question 3: How do pre-service teachers conceptualize (i.e. "think aloud") MKT and how does this conceptualization relate to the development of items for assessing MKT? Preservice teacher conceptualization of MKT was obtained through qualitative data collection. Think aloud interviews with volunteers were conducted following each administration of the items. From those electing to volunteer, a purposeful sample of students was chosen to maximize course level representation and variability in PMKT total scores. Eight think-aloud interviews per preservice course level were solicited. Also, eight additional think-aloud interviews were solicited from a sample of undergraduate psychology students whom volunteered to be administered the PMKT. Interview data were transcribed and then analyzed using NVivo 8 software (QSR International, 2006) for themes related to the conceptualization of MKT and the ability of the PMKT items to assess that knowledge. A priori themes designated by content experts whom developed the items include:

1. Presenting mathematical ideas
2. Responding to students' "why" questions
3. Finding an example to make a specific mathematical point
4. Recognizing what is involved in using a particular representation
5. Linking representations to underlying ideas and to other representations
6. Evaluating the plausibility of students' claims (often quickly)
7. Giving or evaluating mathematical explanations
8. Inspecting equivalencies

Qualitative data obtained from think-aloud interviews were transcribed verbatim. Coding was conducted by two trained coders to minimize experimenter bias. Training was conducted by the principal investigator, who also served as one of the coders. The training process was used to increase coder consistency. Consistency was calculated using Cohen's kappa (1960; Bakeman, 2000). All data was analyzed using NVivo 8.

Topical and analytical coding procedures were used in analyzing the qualitative data. Topical coding was used to identify a priori themes that appeared in the transcripts of the participants. As discussed by Richards (2005), topical coding aided in the preparation of data for interpretive analysis. Analytical coding allowed for the interpretation of the transcribed data by placing meaning to the data. Both the topical and analytical coding was done at the segment unit, where sentences and phrases were analyzed.

There was a two stage process to coding data from the think-aloud interviews. The first stage involved the use of the a priori codes listed above. These codes were used to identify themes related to mathematical knowledge for teaching. Stage two of coding involved inductive coding, which allows codes to emerge from the data (Miles & Huberman, 1994). This process allowed for the addition of new codes or the refinement of a priori codes (Haney, Russell, Gulek, & Fierros, 1998).

Research Question 4: What level of face validity does the new items created to assess mathematical knowledge for teaching have with a pre-service teacher sample? In addition to exploring how the preservice teachers cognitively processed the items, this researcher also explored the face validity of the items. During the cognitive interview the preservice teachers were asked to give their impression of the items. Similar to the cognitive interviews, their responses were transcribed and coded using the a priori codes listed above. Saturation or prevalence of the codes provided an indication of face validity.

Summary

In summary, to validate the use of the PMKT items with pre-service teachers for the purpose of program assessment, the data collected in this study were analyzed using both quantitative and qualitative methods of analysis. Factor analysis was conducted to assess the interrelationships amongst the items. Item analyses were then conducted to address individual item performance. Group differences were then analyzed to address theoretical expectations. Cognitive interviews were then conducted to gather further validity evidence for inferences made using the PMKT items. Those interviews were also used to provide convergent data for assessing item functioning.

CHAPTER FOUR

RESULTS

The purpose of this study was to develop a measure of MKT to be used for preservice program assessment. This scale development process included analyses of item functioning and an examination of construct validity evidence.

Research Questions

The following research questions were addressed to achieve this purpose:

1. How do the PMKT items perform?
    - What factor structure is plausible for the data?
    - How does item difficulty and item discrimination vary?
    - How effective are the distractors?
    - What level of reliability is demonstrated by these new items when used with this pre-service teacher sample?
2. How do groups of pre-service teachers at different academic levels compare on their aggregate scores of the PMKT items?
3. How do pre-service teachers conceptualize (i.e. think about MKT and how does this conceptualization relate to the functioning of items for assessing MKT?
4. What level of face validity do PMTK items have with a pre-service teachers?

Introduction to Results

This chapter details the analyses of a sequential mixed-methods design used to address these questions. First, the process of data cleaning and screening is described. Then factor analyses are discussed in terms of the interrelationships amongst the items. Item performance is addressed through item analysis, including distractor analysis. The factor analysis and item analysis fall within Benson's (1998) structural stage of obtaining construct validity evidence. Second, group differences based on enrollment in IDLS math education courses are explored as suggested by Benson's external stage of construct validity.

The inferences made using Benson's framework are supplemented by the qualitative results. Transcripts from cognitive interviews are analyzed to address how preservice teachers conceptualize MKT. Data from those interviews are then used to explore the face validity of the PMKT items. Finally, data from the cognitive interview are connected to the results of the factor analysis and item analysis in order to provide context to the factor analysis and item analysis results.

Data Cleaning

The data received from all administrations of the 23 PMKT items were screened for univariate and multivariate outliers. These outliers represent extreme scores on the items and atypical scoring patterns. Identifying these extreme cases allowed the researcher to interpret the data accurately. Univariate outliers were screened using a graphical plot of total scores on the PMKT items. Multivariate outliers were screened using a macro for SPSS written by DeCarlo (1997). Analysis of data from all test takers suggests there were neither univariate nor multivariate outliers. The data were then split by course (i.e. Math 107, Math 207, etc.) to further explore outliers. There were no univariate or multivariate outliers in the Math 107 group. There was one multivariate outlier in the Math 207 course. Further examination of the data suggested the individual answered randomly. The case was removed from the data. There were no univariate or multivariate outliers in the ELED 433 group. In addition, there were no multivariate outliers in the non-preservice comparison group. Missing data was not an issue due to the forced response constraint on the computerized answer sheet.

Researchers planned to remove data from all analyses provided by participants who are unmotivated. Unmotivated students were defined as those whom reported an SOS effort subscale score of 10 or less. However, no participants reported a score of 10 or less. Based on the reported SOS effort scores, no participants were removed due to a lack of motivation.

Data Screening

Multicollinearity occurs when variables or items are highly correlated. Univariate multicollinearity was examined by analyzing the tetrachoric correlations of the PMKT items (Table 2). The largest bivariate correlation was .790 for items 9 and 10. This strong correlation may suggest redundancy in the items. Consequently, these items were flagged for possible removal and revision. Univariate normality was also addressed by examining the descriptive statistics. Few variable distributions (items 4, 22, and 23) exceeded suggested cut-offs of an absolute value of 2 for skewness and 7 for kurtosis (Bandalos & Finney, 2010). Univariate descriptive statistics in Table 2 indicate that the data are relatively normally distributed.

Confirmatory Factor Analysis

The first research question of this study focused on item functioning. Initial steps in addressing the functioning of the items included factor analysis. This technique allowed the researcher to assess the structure of the item interrcorrelations, as suggested Benson's (1998) strong program of construct validation. By first conducting the factor analysis the researcher was able to create a foundation for interpreting discrimination values obtained later in the item analysis (CITE).

A confirmatory factor analysis (CFA) was conducted using the Mplus software (Muthén & Muthén, 2007). A unidimensional model was fit to the data. This unidimensional model was analyzed to obtain evidence that the PMKT items are measuring one construct, specialized content knowledge. Identifying adequate model fit would inform the item development process.

Due to the binary nature of the data, a tetrachoric correlation matrix (Table 2) with robust weighted least squares (WLMSV) estimation method was employed (Muthén, 2009). Conducting a CFA with binary data typically requires large sample sizes and a cautious concern for non-normality and biased standard errors (Muthén & Kaplan, 1992). This is due to the model being based on the underlying continuous, bivariate, non-normal distribution. However, the WLMSV estimation procedure lessens the sample size requirements, and is more robust with respect to issues that arise when data are non-normal. It also provides accurate test statistics, parameter estimates, and standard errors under a variety of conditions (Flora and Curran, 2004). Moreover, the WLMSV does not require the process of inverting the weight matrix, and hence the problem of non-positive definite matrices caused by skewed items is avoided all together (Brown, 2006).

Table 2.
Tetrachoric Correlations for PMKT items (N=396)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	---																						
2	0.50	---																					
3	0.05	0.02	---																				
4	-0.19	0.13	0.09	---																			
5	0.29	0.26	0.16	-0.03	---																		
6	0.21	0.29	0.10	0.09	0.20	---																	
7	0.20	0.39	0.21	0.22	0.21	0.22	---																
8	0.31	0.17	0.28	0.07	0.24	0.21	0.34	---															
9	0.10	0.18	0.36	-0.01	0.20	0.20	0.21	0.17	---														
10	0.13	0.16	0.17	0.01	0.16	0.07	0.22	0.05	0.79	---													
11	0.19	0.18	0.02	-0.02	0.08	0.11	0.13	0.24	0.22	0.17	---												
12	0.18	0.00	0.04	-0.04	0.17	0.15	0.24	0.20	0.20	0.25	0.35	---											
13	0.20	0.14	0.11	0.09	0.31	0.12	0.08	0.19	0.31	0.11	0.09	-0.11	---										
14	0.21	0.19	0.13	0.03	0.16	0.35	0.29	0.08	0.32	0.25	0.34	-0.02	0.40	---									
15	0.13	-0.02	0.08	-0.04	0.03	0.04	-0.03	0.03	0.08	0.09	0.12	0.05	-0.01	0.18	---								
16	-0.13	-0.15	0.08	0.38	-0.01	0.07	-0.12	0.04	-0.04	-0.04	-0.10	0.02	0.01	0.12	0.30	---							
17	0.13	0.14	0.09	0.06	0.10	0.18	0.12	0.14	0.27	0.02	0.27	0.11	0.24	0.37	0.15	0.17	---						
18	0.09	-0.02	-0.20	0.06	0.09	-0.06	0.03	0.07	-0.12	-0.16	-0.06	-0.01	-0.09	-0.12	-0.02	0.07	-0.07	---					
19	0.16	0.18	-0.01	-0.01	0.01	0.21	0.04	0.01	0.07	0.09	0.20	0.09	0.10	0.11	-0.04	0.03	0.08	-0.15	---				
20	0.06	0.06	0.13	0.02	0.19	0.17	0.12	0.21	0.22	0.16	0.19	0.02	0.12	0.07	-0.03	0.13	0.15	0.02	0.13	---			
21	0.18	0.25	0.21	0.08	0.25	0.13	0.15	0.20	0.30	0.17	0.22	0.02	0.31	0.47	0.08	0.02	0.16	-0.02	0.10	0.08	---		
22	0.22	0.25	0.06	0.04	0.22	0.38	0.23	0.30	0.35	0.24	0.31	0.10	0.36	0.12	-0.04	-0.10	0.05	0.01	0.15	0.16	0.28	---	
23	0.26	0.15	-0.03	-0.34	0.13	0.07	0.09	0.07	0.01	0.13	0.12	0.11	0.02	-0.03	0.01	-0.04	-0.03	0.08	0.11	0.14	-0.12	0.15	----
M	0.59	0.75	0.55	0.90	0.16	0.84	0.88	0.60	0.60	0.58	0.48	0.44	0.74	0.84	0.31	0.22	0.71	0.44	0.35	0.23	0.78	0.86	0.11
SD	0.49	0.44	0.50	0.32	0.37	0.37	0.32	0.49	0.49	0.49	0.50	0.50	0.44	0.37	0.46	0.42	0.46	0.50	0.48	0.42	0.42	0.35	0.32
Skew	-0.36	-1.13	-0.21	-2.44	1.82	-1.87	-2.41	-0.41	-0.39	-0.34	0.09	0.26	-1.11	-1.82	0.84	1.36	-0.90	0.26	0.63	1.27	-1.32	-2.07	2.44
Kurt	-1.88	-0.73	-1.96	3.99	1.32	1.51	3.80	-1.84	-1.86	-1.89	-2.00	-1.95	-0.77	1.32	-1.31	-0.15	-1.20	-1.95	-1.62	-0.38	-0.25	2.28	3.99

Model fit was evaluated using the χ^2 as well as four other fit indices discussed by Brown (2006). The chi square statistic provided an assessment of absolute fit of the data. The comparative fit index (CFI), Tucker-Lewis index (TLI), Root mean square error of approximation (RMSEA), and Weighted root-mean-square residual (WRMR) each provide an index of relative fit by comparing the model implied and covariance matrices. Yu & Muthén (2002) recommended that the following cut offs for determining good fit for each of the indices: CFI greater than .96, TLI greater than .95, RMSEA at or below .05, and WRMR less than or equal to 1.0. The fit criteria for each index will be used to examine the fit of the model (Hu & Bentler, 1999).

The model fit data is presented in Table 3. The unidimensional model failed to meet the cut offs for all indices with the exception of RMSEA. Table 4 includes the path coefficients for the items. The completely standardized coefficients for most items on the scale were low (i.e., less than .60). Items 9 and 10 were the only items with relatively high path coefficients (i.e., .84 and .70). The variance explained by the factor (i.e. R^2) was low for most items. Specifically, items 3, 4, 12, 15, 16, 17, 18, 19, 20, and 23 had R^2 values less than .15.

Table 3.
Fit Statistics for Unidimensional Model

Model	χ^2	df	WRMR	RMSEA	CFI	TLI
23-item, one-factor	218.959*	122	1.125	0.045	0.794	0.816

*P= <.0001

Table 4.
Unstandardized (Standardized) Parameter Estimates and Variance Explained

Items	Path Coefficients	R^2 Value
1	1.00 (.41)	0.17
2	1.05 (.44)	0.19
3	0.75 (.31)	0.10
4	0.11 (.05)	0.00
5	0.94 (.39)	0.15
6	0.96 (.40)	0.16
7	1.05 (.44)	0.19
8	0.94 (.39)	0.15
9	2.02 (.84)	0.70
10	1.69 (.70)	0.50
11	1.00 (.42)	0.17
12	0.67 (.28)	0.08
13	0.98 (.41)	0.17
14	1.28 (.53)	0.28
15	0.32 (.13)	0.02
16	-0.03 (-.01)	0.00
17	0.82 (.34)	0.12
18	-0.28 (-.12)	0.01
19	0.50 (.21)	0.04
20	0.68 (.28)	0.08
21	1.11 (.46)	0.21
22	1.20 (.50)	0.25
23	0.34 (.14)	0.02

Exploratory Factor Analysis

Due to the poor fit of the CFA an exploratory factor analysis was used to uncover underlying factors that drive the responses to these items. This procedure was conducted by analyzing the sample tetrachoric correlations in SPSS. Due to non-convergence of the factor solution, principal axis factoring was abandoned and replaced with unweighted least squares estimation (ULS). ULS estimation is robust to data issues related non-normal data (Nunnally & Bernstein, 1994). For this reason the ULS estimation procedure was likely to converge when other methods do not.

Two statistics were used to quantify the level of correlations among the PMKT items. Bartlett's Test of Sphericity (to test the null hypothesis that the correlation matrix is an identity matrix) and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (tests the size of the partial correlations to determine if factor analysis is inappropriate; see Kaiser, 1970) were used to determine the appropriateness of factor analytic procedures. The KMO value of .468 suggests that the correlation matrix is unacceptable for conducting factor analysis (Kaiser, 1974). However, Bartlett's test of sphericity (p<.0001) suggests that the correlation matrix deviates enough from the identity matrix to conduct factor analysis (Snedecor & Cochran, 1983). Given the exploratory nature of this study, the factor analysis was conducted despite the conflicting statistics regarding the level of correlation among the variables.

There are several rotation methods that could be used to identify the structure of the factor solutions. These methods exist within two categories, orthogonal and oblique. Unlike orthogonal rotation, oblique rotation takes into account the relationships that exist among your factors. This process allows for factors to correlate, resulting in the ability to accurately interpret relationships between factors and variables. In this study *direct oblimin*, a form of oblique rotation was used. Since MKT is thought to consist of several inter-related knowledge domains multiple factors in this data set were expected to be correlated. The direct oblimin rotation procedure was used to achieve simple structure by reducing the number of salient cross loadings.

As suggested in Bandalos and Finney (2010), deciding on the number of factors to extract should involve several solutions. Here the eigenvalue greater than 1 rule (K1), a scree plot, and parallel analysis were used to statistically determine the number of plausible factors. The quality of each factor was determined by analyzing communalities, pattern coefficients and structure coefficients. Communalities represent the proportion of variance that is shared amongst items in a factor solution. Items with communalities less than .20 were considered for removal. Pattern and structure matrices assisted in the interpretation of factors by providing estimates of the relationship between items and factors. The structure coefficients represent the direct and indirect relationship between factors and items. The pattern coefficients illustrated the unique relationship between the variables and the factors after removing its relationship with other variables. As suggested in Gorsuch (1983), salient loadings can be used to attribute items to factors. A common criterion for determining saliency is an absolute value of .3. Additionally, cross-loadings will be used to determine the salience of factors. Lastly, a theoretical consideration of each factor pattern was addressed.

The K1 rule suggested that 9 factors could be retained. Initial eigenvalue for each factor were 4.30 (factor 1), 1.95 (factor 2), 1.62 (factor 3), 1.46 (factor 4), 1.43 (factor 5), 1.27 (factor 6), 1.20 (factor

7), 1.08 (factor 8), and 1.06 (factor 9). The scree plot (Figure 3) suggested a three factor solution should be retained. Use of the mean values as a point of reference for the parallel analysis procedure indicated that a three factor solution was plausible (Table 5).

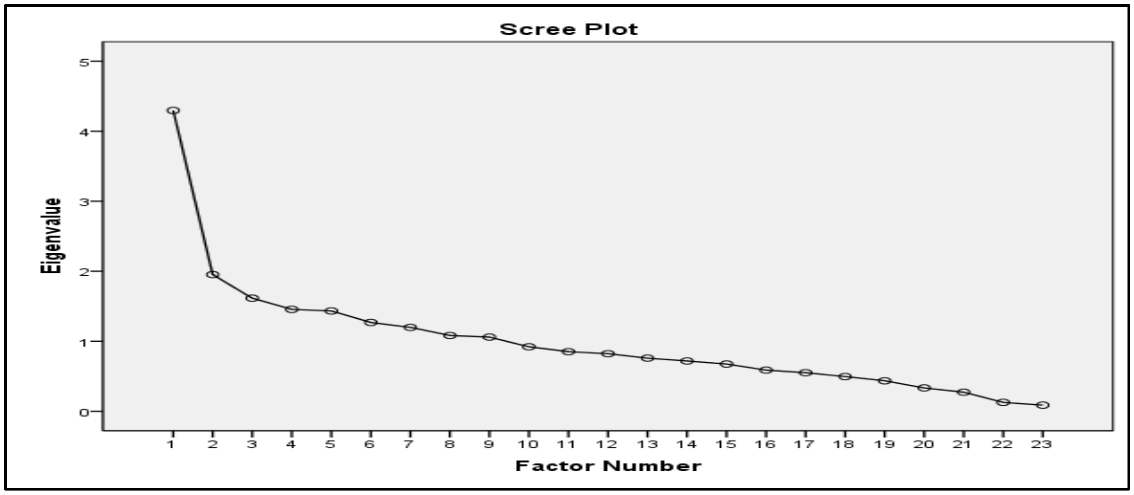


Figure 3.*Eigenvalue scree plot*

Table 5.
Raw Data Eigenvalues, & Mean & Percentile Random Data Eigenvalues

Root	Raw Data	Means	Percentile
1	1.90	1.45	1.52
2	1.47	1.38	1.42
3	1.37	1.33	1.37
4	1.28	1.28	1.32
5	1.21	1.23	1.27
6	1.14	1.19	1.23
7	1.11	1.15	1.19
8	1.08	1.12	1.15
9	1.06	1.08	1.11
10	1.03	1.05	1.08
11	0.99	1.02	1.04
12	0.96	0.98	1.01
13	0.91	0.95	0.98
14	0.89	0.92	0.94
15	0.87	0.89	0.92
16	0.84	0.86	0.89
17	0.79	0.83	0.86
18	0.77	0.80	0.83
19	0.75	0.77	0.80
20	0.68	0.74	0.77
21	0.67	0.70	0.74
22	0.64	0.66	0.69
23	0.59	0.61	0.65

Theoretically, a one factor solution was most plausible given that all items were written to measure specialized content knowledge of mathematics. However, this domain of MKT was operationalized by 6 learning objectives (Table 1). In all but two cases, each item was mapped to Objective 1 and one or two other objectives. For these reasons, a 1 factor, 3 factor, 4 factor, 5 factor, and 6 factor solution were initially considered. However, retaining a number of factors beyond 3 was not supported by the scree plot and parallel analysis. Therefore, given the theoretical and the empirical data, a one factor and a three factor solution were analyzed.

After extracting three factors 26.43% of the cumulative variance in the variables was accounted for. Eigenvalues for each of the factors were 3.67 (factor 1), 1.29 (factor 2), and 1.12 (factor 3). The amount of variance in each variable that is shared with all other variables is illustrated by the communalities (Table 6). The communalities were low to moderate. Items 3, 11, 12, 15, 17, 18, 19, and 20 had communalities less than .20. A closer examination of the communalities also indicated that a Heywood Case exists within the data. This implied that some unique factor lacks the variance needed to identify a solution. This could have been caused by the extraction of too many or too few factors, small data sets, or model misspecification. Examination of the pattern and structure loadings (Tables 7 and 8) indicated that items 20 and 19 appeared to load onto factor 1, but their loadings were low. Items 9 and 10 appeared to form their own factor, while items 3 and 18 had low loadings on that same factor. Items 15 and12 did not appear to load onto any factor. Factors 2 and 3 did not make conceptual sense. Consequently, the three factor solution was abandoned.

Table 6.
Communalities after 3 factor extraction

Items	Communalities
1	0.47
2	0.34
3	0.14
4	0.26
5	0.21
6	0.23
7	0.22
8	0.29
9	1.00[b]
10	0.66
11	0.19
12	0.10
13	0.23
14	0.40
15	0.04
16	0.20
17	0.19
18	0.05
19	0.06
20	0.08
21	0.29
22	0.29
23	0.25

[b] Heywood case

Table 7.
Pattern Coefficients for the PMKT Items

	Factor		
Item	1	2	3
1	**0.69**	0.19	-0.25
2	**0.63**	0.15	-0.06
8	**0.51**	0.08	0.05
6	**0.49**	0.07	0.12
22	**0.49**	-0.10	-0.08
7	**0.47**	-0.12	0.02
5	**0.47**	0.16	-0.03
14	**0.45**	-0.11	0.34
21	**0.41**	-0.08	0.26
11	**0.39**	-0.09	-0.04
13	**0.39**	-0.06	0.21
17	**0.31**	-0.04	0.28
19	**0.25**	0.01	-0.03
20	**0.23**	-0.10	0.02
12	0.19	-0.16	-0.19
9	0.03	**-1.00**	-0.13
10	-0.02	**-0.83**	-0.28
3	0.11	**-0.25**	0.14
18	0.08	**0.23**	-0.03
4	0.06	0.09	**0.52**
16	-0.03	0.06	**0.46**
23	0.24	0.01	**-0.44**
15	0.06	-0.05	0.15

Table 8.
Structure Coefficients for the PMKT Items

	Factor		
Item	1	2	3
1	0.59	-0.05	-0.28
2	0.56	-0.12	-0.09
22	0.53	-0.29	-0.05
14	0.50	-0.40	0.38
7	0.47	-0.23	0.04
8	0.47	-0.16	0.04
6	0.47	-0.19	0.12
5	0.46	-0.18	-0.02
21	0.45	-0.33	0.29
11	0.43	-0.26	-0.01
13	0.42	-0.28	0.23
17	0.33	-0.25	0.30
20	0.27	-0.21	0.05
12	0.26	-0.19	-0.14
19	0.24	-0.09	-0.03
9	0.48	-0.99	0.13
10	0.34	-0.75	-0.06
3	0.22	-0.34	0.21
18	-0.02	0.20	-0.09
4	0.03	-0.07	0.50
16	-0.05	-0.05	0.44
23	0.23	0.02	-0.44
15	0.09	-0.12	0.16

After removal of 12 items due to low communalities and ambiguous factors, a one factor model was analyzed. The KMO value of .607 indicated mediocre factorability. Again, Bartlett's test of sphericity (p<.001) suggested that the correlation matrix deviates enough from an identity matrix to conduct factor analysis. The K1 rule suggested two factors should be retained. However, the scree plot and the parallel analysis suggested that a 1 factor solution be retained. This one factor solution accounted for 24.87% of the variance in the items. The eigenvalue for factor 1 was 2.74. The communalities for four of the 11 items were below .20 (Table 9). A review of the factor pattern matrix indicated that all items had loadings greater than .30 on factor 1 (Table 10).

Table 9.
Communalities

Items	Communalities
1	0.19
2	0.26
5	0.20
6	0.19
7	0.19
9	0.38
10	0.25
13	0.23
14	0.31
21	0.26
22	0.28

Table 10.
Factor Loadings

Items	Factor 1
9	0.62
14	0.56
22	0.53
21	0.51
2	0.51
10	0.50
13	0.48
5	0.44
1	0.44
7	0.44
6	0.43

Naming Factors

Notwithstanding the low communalities, the 1 factor solution was interpreted. This factor consists of 11 items. With each of these items there appears to be an underlying emphasis on evaluating mathematical arguments. Each of these items was previously mapped to a learning objective concerning the evaluation of a K-8 student's mathematical work or arguments to determine if the ideas presented are valid. For example, item 1 asks the examinee to determine the veracity of three methods for subtracting large numbers. These remaining 11 items account for 58% of the original items written to assess this objective.

Item Analysis

To further address research question 1, item analyses were used to determine the functioning of each PMKT item. This process allowed the investigator to identify how each item may be improved. The functioning of each item was determined by the item difficulty (*p*-value) and discrimination, as well as a distractor analysis. Item responses from all 396 preservice teachers were analyzed. Responses from a comparison group of non-preservice teachers were also analyzed.

Table 12 presents the item analysis of the psychometric properties of the PMKT items. The categorization of data is as follows: All participants, all preservice teachers (Math 107, Math 207, and ELED 433), Upper Level preservice teachers (Math 207 and ELED 433), ELED 433, Math 207, Math 107, and Comparison Group (Comp). In the sample of all preservice teachers, item difficulty (*p)* ranged from .11 to .89. Four questions were below .30 indicating harder questions (Items: 5, 16, 20, and 23) and five were above .80 indicating easier question (Items: 4, 6, 7, 14, and 22). The p-value for Item 5 was .31 for ELED 433 group, and less than .20 for all other groups. This was the same trend for item 23. The *p* value for item 4 was the highest. This item was relatively more difficult for ELED 433 than MATH 207 and MATH 107. MATH 207 performed the same or relatively better than ELED 433 on the other items flagged as easy.

For all preservice teachers the item discrimination ranged from -.06 to .39. Items 3, 4, 12, 15, 16, 18, 19, 20, 23 were flagged for low discrimination values (r_{pbis}<.20). For all preservice teachers, incorrectly responding to item 18 (r_{pbis}=-.06) was associated with higher total scores. Items 12 and 15 produced better discrimination values in the ELED 433 subsample. Overall, the discrimination values improved when analyzed with the ELED 433 sample. The discrimination values for the revised 11-item scale were similar to those obtained for the same items in the 23-item scale (Table 11).

Table 11.
Item Discrimination Comparison

	23-item PMKT	11-item PMKT
Item	r_{pbis}	r_{pbis}
1	.28	.27
2	.26	.31
5	.22	.21
6	.24	.22
7	.23	.22
9	.39	.41
10	.28	.32
13	.22	.27
14	.30	.31
21	.27	.29
22	.27	.29

The reliability estimate using all items for the preservice teacher sample was .61. Coefficient alpha reached .61 and .68 when estimated with the Upper level preservice subsample and the ELED 433 subsample, respectively. Items were removed with difficulties less than .30, discrimination less than .20, and low communalities or high cross-loadings in the factor analysis. Using the remaining 11-item scale, alpha = .63 for all preservice teachers. Alpha reached .66 and .73 with the Upper level preservice subsample and the ELED 433 subsample, respectively.

Table 12.

Item Difficulty and Discrimination

			All		Upper Level								Non-	
	All		Preservice		Preservice		ELED		MATH		MATH		Preservice	
	Participants		Teachers		Teachers		433		207		107		Teachers	
Item	*p*	r_{pbis}	*p*	r_{pbis}	*p*	r_{pbis}	*p*	r_{pbis}	*p*	r_{pbis}	*p*	r_{pbis}	*p*	r_{pbis}
1	.56	.30	.59	.28	.73	.25	.91	.12	.63	.28	.50	.25	.52	.31
2	.70	.25	.75	.26	.81	.24	.85	.31	.78	.18	.71	.24	.63	.22
3*	.53	.14	.55	.17	.61	.07	.55	.02	.65	.15	.52	.20	.50	.10
4*	.89	.05	.89	.04	.86	-.01	.75	-.02	.92	.11	.90	.10	.89	.08
5	.16	.15	.16	.22	.23	.17	.31	.30	.18	.04	.12	.22	.16	.04
6	.84	.19	.84	.24	.90	.19	.87	.30	.92	.13	.80	.23	.84	.11
7	.88	.21	.88	.23	.96	.14	.96	.33	.96	.02	.84	.23	.88	.19
8*	.59	.27	.60	.30	.69	.21	.83	.33	.61	.12	.55	.30	.57	.23
9	.58	.38	.60	.39	.60	.44	.65	.51	.57	.39	.60	.38	.55	.35
10	.58	.29	.58	.28	.61	.34	.64	.41	.59	.28	.57	.25	.58	.32
11*	.51	.28	.48	.28	.49	.27	.55	.29	.46	.23	.47	.30	.55	.30
12*	.44	.14	.44	.18	.48	.19	.49	.30	.48	.12	.41	.15	.43	.09
13	.74	.21	.74	.22	.80	.19	.84	.23	.78	.15	.70	.22	.73	.20
14	.81	.30	.84	.30	.88	.37	.85	.39	.90	.40	.81	.24	.77	.30
15*	.26	.11	.31	.10	.27	.21	.27	.35	.26	.13	.33	.07	.18	.10
16*	.21	.06	.22	.04	.18	.05	.16	.12	.20	.01	.24	.06	.19	.10
17*	.72	.18	.70	.22	.71	.23	.78	.18	.66	.24	.70	.23	.74	.13
18*	.41	-.05	.44	-.06	.44	.00	.49	.13	.40	-.11	.44	-.10	.37	-.04
19*	.35	.09	.35	.12	.40	.16	.45	.07	.37	.21	.32	.08	.36	.05
20*	.26	.16	.23	.18	.30	.11	.29	.05	.30	.16	.20	.20	.30	.15
21	.77	.23	.78	.27	.82	.42	.82	.61	.82	.30	.75	.18	.77	.18
22	.87	.25	.86	.27	.90	.33	.89	.36	.91	.34	.83	.22	.89	.23
23*	.09	.08	.11	.07	.20	.07	.36	.14	.10	-.10	.06	-.01	.07	.08
N	665		396		147		55		92		249		269	
Min	3		3		5		5		5		3		5	
Max	21		21		21		21		21		21		21	
Mean	12.74		12.94		13.88		14.58		13.46		12.39		12.45	
SD	3.17		3.25		3.16		3.44		2.91		3.18		3.03	
Alpha	.59		.61		.61		.68		.54		.58		.55	

*Removed during item analysis

Distractor Analysis

Table 13 provides an example of information data obtained from the distractor analysis. There were several response options that were not often chosen by the lowest scoring 27% of preservice teachers. For example, distractor B in item 4 did not provide much information for discriminating between high and low scorers. This is the same case for distractors A, C, and D for item 21. These distractors should be revised or removed. Reducing the number of low performing distractors can increase item difficulty (Cizek & Day, 1994).

Table 13.
Distractor Analysis (Hi N=126, Low N=129)

Item	*p* value	r_{pbis}	% of Responses per Option for High 27% and Low 27% Scorers							
4	.89	.04		A*	B	C				
			Low 27%	84	1	16				
			Hi 27%	91	0	9				
21	.78	.27		A	B	C	D	E	F*	G
			Low 27%	2	19	1	1	4	57	16
			Hi 27%	0	2	1	0	0	95	2

Between Groups Analysis

Group differentiation analyses were examined to address research question two. These analyses allowed the researcher to address the external stage of a strong program of construct validity. After the removal of items based on the factor analysis and item analysis, a between groups ANCOVA was conducted to determine whether subsamples of participants differed on their 11-item PMKT totals after controlling for SAT (Table 14). Preservice teachers with missing SAT MATH scores were removed from the analysis. There were 225 MATH 107 participants, 73 MATH 207 participants, and 53 ELED participants for whom there were SAT MATH scores available. Though the sample sizes of the grouping variable differed, Levene's test of equal error variances indicated that the homogeneity of variances assumption was not violated (p=.772). The covariate, SAT MATH, was significantly related to PMKT *F* (1, 347) =125.21, p<.001, r=.48. There was also significant effect of preservice course level on PMKT after controlling for SAT MATH, *F* (2,347) =12.69, p<.001. No significant interaction between SAT MATH and preservice course level was found. Planned contrast (Table 15) revealed that the mean on the PMKT items for MATH 207 is significantly greater than the mean on the PMKT items for MATH 107, *t*(347)= 3.769, p<.001, *r*=.20. Also, the PMKT mean for ELED 433 was significantly greater than the mean score for MATH 107, *t*(347) =4.684, p<.001, *r*=.24. There was no significant difference between ELED 433 and MATH 207, *t*(347) = -1.645, p=.102.

Table 14.
Analysis of Covariance Summary

Source	Sum of Squares	df	Mean Square	F	Partial Eta Squared
SAT Math	364.14	1	364.13	110.40**	.23
Preservice Course	83.57	2	41.79	12.67**	.05
Error	1144.56	347	3.298		

**p < 0.01

Table 15
Bonferroni Comparison for Preservice Course Level

Comparisons	Mean Score Difference	Std. Error	Effect Size	95% CI Lower Bound	95% CI Upper Bound
ELED 433 vs. MATH 207	.21	0.33	.09	-1.00	0.58
ELED 433 vs. MATH 107	1.13**	0.28	.24	0.46	1.80
MATH 207 vs. MATH 107	.92**	0.25	.20	0.33	1.51

** p < 0.01

Quantitative Results Summary

Summary of Research Question 1

Factor analysis and item analysis were used to address item functioning. The factor analyses were conducted first to address the dimensionality of the data. The confirmatory factor analysis indicated a lack of unidimensionality for the 23 PMKT items. However, exploratory factor analysis indicated that the initial 23 items could plausibly be pared down to an 11 item essentially unidimensional scale. Item analysis was conducted to further identify item functioning across participants. Several of the same items flagged for questionable performance in the factor analysis were again flagged in the item analysis. The items flagged in the item analysis were removed, leaving 14 items for estimating reliability. The difference between these 14 items and the revised 11 item scale obtained in factor analysis were items 8, 11, and 17. Given the sample of preservice teachers (N=396), the reliability estimate for the 11 item scale and the 14 item scale was .63 and .66 respectively. Despite this slight improvement in reliability with the 14 item scale, the 11 item scale provided the best structural validity evidence for measuring the MKT construct.

The estimate of reliability, similar to other item statistics, improved as the preservice teacher subgroups changed from MATH 107 to ELED 433. This was the case for the 23 item scale, as well as the revised scales. For instance, when using the 11 item scale the reliability estimates were .58 for MATH 107, .61 for MATH 207, and .70 for ELED 433. When combining the MATH 207 group and the ELED 433 group the reliability estimate was .65. This indicated that the internal consistency of the items increased as the preservice teachers' course level increased.

Summary of Research Question 2

An analysis of group differences was needed to determine how well students at each preservice course level performed on the items. An ANCOVA was used to control for the variance in PMKT scores associated with general math ability as measured by SAT Math scores. Using the set of 11 PMKT items significant differences were exhibited between ELED 433 and MATH 107, and between MATH 207 and MATH 107. There was no significant difference between PMKT total score of the ELED 433 and MATH 207 groups, though there were several items on which the two groups differed. Determining which misconceptions are plausible to explain the difference in mean score on those items could provide insight in how to improve items that should but are unable to differentiate between these two levels of PMKT.

Think-aloud Analysis

A priori coding results. Following the calibration of coders on sample transcripts of cognitive interviews, each of the two coders proceeded with independently coding transcripts. Each transcript included verbal responses of participants thinking aloud about the PMKT items, as well as their responses to retrospective questions. A total of 21 transcripts were coded. The coding process included the coding of 5 transcripts from the MATH 207 group and 5 from the ELED 433 group. The coding process also included the coding of 7 transcripts from the MATH 107 and 4 transcripts from the comparison group.

In stage one, a priori codes were used to initially identify how the preservice teachers and comparison group conceptualized the items. The transcribed response to each item was reviewed and coded using a priori codes. In stage two emergent themes related to their thought process in responding to the items were developed. Also, themes related to item functioning (e.g. ambiguous wording) were noted.

Raw percent agreement between the coders was calculated to address reliability or consistency of codes assigned to the qualitative data. Raw percent agreement does not consider chance agreement. For this reason, percent agreement beyond chance was also calculated using Cohen’s Kappa (Cohen, 1960). Though Kappa is a more robust estimate of reliability than its raw counterpart, it does have some limitations. For example, chance agreement is determined by calculating an expected level of agreement which can be biased and produce overly conservative estimates of reliability. For this reason, inter-coder agreement was assessed using both raw percent agreement and Kappa. Table 16 provides agreement statistics for the a priori codes.

Table 16.
Inter-coder Agreement Results

	% Agreement	Kappa	# of coded units
Presenting mathematical ideas	0.98	0.88	5
Responding to students' "why" questions	0.90	0.79	22
Finding examples to make a mathematical point	0.83	0.80	21
Recognizing what is involved in using a particular representation	0.76	0.66	15
Linking representations to underlying ideas and to other representations	0.75	0.69	17
Evaluating the plausibility of students' claims (often quickly)	0.86	0.62	49
Giving or evaluating mathematical explanations	0.72	0.60	73
Inspecting equivalencies	0.62	0.56	5
Overall	0.69	0.62	135

The overall agreement for the coders was .62. Based on the rules of thumb for interpreting kappa, the overall agreement was "substantial" (Landis & Koch, 1977). However, there was "outstanding" agreement for *Presenting mathematical ideas*. Most codes were within the "substantial" range of agreement (i.e. .61-.80). Consensus between coders was obtained through discussion.

Most disagreements between the coders could be categorized as a difficulty in managing multiple codes for the same segments of text. This was due to some overlap between similar a priori codes. In each case of disagreement, consensus was made by deciding to utilize the overlapping coded text instead of removing the coded unit of transcription. For example, Coder 1 assigned *evaluating the plausibility of students' claims* and *giving or evaluating mathematical explanations* to several of the same units of transcription. In each of these cases, Coder 2 only assigned *giving or evaluating mathematical explanations* to those same units. During the process of consensus, it was decided that the two codes possess similar characteristics, but they were not the same. It was further decided that the units coded by Coder 1 were most plausible.

Once consensus regarding a priori codes at the unit level was obtained, the saturation of codes across all transcripts was then analyzed. The transcribed verbal responses indicated that *giving or evaluating mathematical explanation* was most illustrative of the of the thought process used by the participants when responding to the items and responding to the retrospective interview questions. This code was utilized on three times as many units as the next most saturated code. Though mostly used by preservice teachers in the MATH 207 and ELED 433 group, this code was also prevalent in the MATH 107 and Comparison groups. For example, items 1 and 6 were coded as *giving or evaluating mathematical explanations* by all preservice teachers whom participated in the think-aloud interviews. *Inspecting equivalencies and* presenting *mathematical ideas* were the least coded verbal responses to either the items or the retrospective questions. These codes were most prevalent in the ELED 433 and MATH 207 group. These codes were rarely illustrative of the thought process verbalized by the MATH 107 and comparison group. Table 17 provides further detail regarding the number of codes utilized on responses to the PMKT items and those utilized on responses to retrospective questions.

Table 17.
Codes by Prompt

Codes	# of Codes across PMKT items	# of Codes across Retrospective Questions
Presenting mathematical ideas	0	5
Responding to students' "why" questions	5	17
Finding examples to make a mathematical point	6	15
Recognizing what is involved in using a particular representation	10	5
Linking representations to underlying ideas and to other representations	9	8
Evaluating the plausibility of students' claims (often quickly)	31	18
Giving or evaluating mathematical explanations	44	29
Inspecting equivalencies	5	0

A priori codes indicated that preservice teacher responses to these items were most associated with evaluating mathematical explanations and evaluating the plausibility of students' claims. Other codes were more prevalent in the qualitative data obtained from the upper level preservice teacher participants (i.e. Math 207 and ELED 433) than from the MATH 107 group and the comparison group. This difference is illustrated in the response to Item 1 by a high scoring participant from ELED 433 and a low scoring participant in the MATH 107 group. This difference was not as clear between preservice teachers in MATH 207 and ELED 433. Table 18 provides an illustration of those differences.

Table 18.
Verbal Report Excerpt for Retrospective Question 1: What type of knowledge or skill is required to answer the questions you have been presented with?

ELED 433 Participant	MATH 107 Participant
P = I guess like, it is like more than just basic math skills, like more than just knowing like 5 +2 is 7 like to is a lot more than that. So, I guess you have to have [MATH] 107 and all that stuff. But, I guess it is more like critical thinking and like analyzing and I don't know we always talk about like you have to really try to figure out what the student is thinking. And, like analyze their work instead of just looking for the right or wrong answer, so, I guess you still have to learn how to analyze stuff…If my student would like draw a picture…I would have to decide if the picture matches what their thinking is and talk with them about their thinking to really like know if they understand the math concept or not.	P = um, division, multiplication, word problem sense. Like how to figure out the, what to, what kind of thing you would use in order to get the answer. I = what do you mean? P = like, um when you are reading a problem, you have to decide whether its division or multiplication, um, so just the problems we did probably knowledge of multiplication, dividing, and fractions. I = ok. So just your ability to do the procedure is enough? P = um just how to do it

P = the participant, I = the investigator

As identified by the a priori codes, the ELED 433 participant addressed *giving or evaluating mathematical explanations*, *evaluating the plausibility of students' claims, and linking representations to underlying ideas and to other representations*. The MATH 107 participant's response did not warrant the use of the a priori codes. Instead this participant's response indicated a fundamental reliance on the procedural knowledge. Though this basic knowledge of mathematical operations is necessary for teaching, it does not reflect the knowledge that should be elicited by the items, nor does it reflect the knowledge that was made explicit by the learning objectives of the program. The differences between these two participants in their response to retrospective question 1 and their scores on the PMKT items (i.e., a difference of 10 points) provides a plausible indication of different levels of mathematical knowledge for teaching.

In another case, a low scoring ELED 433 participant was compared to a high scoring participant in the MATH 107 group. In a similar fashion, the MATH 107 participant championed procedural knowledge as the main requirement to answer the PMKT items. However, unlike the previous MATH 107 participant this participant also mentioned the use of critical thinking. The low scoring (i.e. 9) ELED 433 participant did mention procedural knowledge, but also elaborated on critical thinking. The ELED participant also focused on how children work through math problems. Two themes were extrapolated from this comparison: 1) In spite of their scores, ELED 433 and MATH 207 participants demonstrated a more intricate understanding of the knowledge the items were intended to measure; 2) ELED 433 and MATH 207 were more likely to mention an understanding children's responses than MATH 107 and the comparison group. In fact, no participant in the comparison group mentioned children in their response to retrospective question 1.

Retrospective questions 3 and 5 were used to further explore the face validity of the items. Question 3 addressed whether the skill required to answer the items was more than basic math operation skills (i.e., procedural knowledge). All cognitive interview participants, including those from the comparison group, noted that correctly responding to all items required more than procedural mathematics knowledge. The response provided by some participants contradicted their early responses to retrospective question 1 regarding the preponderance of procedural knowledge elicited by the items. The articulation of other types of knowledge or skills was more apparent in responses to retrospective question 3. However, the complexity of this articulation differed within and across subgroups of participants. Table 19 provides a comparison of responses from two MATH 207 participants whom obtained similar total scores on the PMKT items. In this comparison, Participant A's response was coded under *giving or evaluating mathematical explanations* and evaluating *the plausibility of students' claims*. The response from Participant B was not coded under any of the a priori codes. However, it is clear from Participant B's response that there is a conceptualization of MKT that extends beyond procedural knowledge when responding to these items.

Table 19.

Verbal Report Excerpt for Retrospective Question 3: Some people would say that skill required to answer these questions are nothing more than basic math operation skill. What would you tell those people?

MATH 207 Participant A	MATH 207 Participant B
P = You had to like know how the kids think about math problems. Like how many different kind of ways to respond to your math problems. Wait. You have to see what they are doing and interpret it. If you give them a problem, you need to know how many ways the problem can be answered and being able to see what they are doing if they are doing it wrong.	P = knowing how to do math problems is helpful but it's not everything. There's [more] to understand about math. I = okay. um do you have an example of what more you need to understand? P = like more skill. I mean you have to know about math.

P = the participant, I = the investigator

Retrospective question 5 further addressed the conceptualization of the knowledge required to answer the PMKT items. The question asked whether higher scores on the PMKT items were associated with greater competence in teaching mathematics. The responses to this question were varied amongst upper level preservice participants, but were fairly consistent within and across the MATH 107 and comparison group. Both of the latter groups provided some indication that correct responses to these items could indicate greater ability to teach mathematics. Participants in the upper level preservice groups, on the other hand, differed on whether answers to these items would reflect their competence to teach mathematics. The rationale most prevalent for why these items would not be representative competence was that the PMKT items were too difficult. One participant cited difficulty in responding to complex word problems. Table 20 provides a response from a MATH 207 student regarding the relationship between MKT and these items. This participant correctly answered 3 more items than the mean number of items correctly answered by the MATH 207 group. This participant's response was not coded using any of the a priori codes. It is important to note that this same participant's response to retrospective question 1 was coded using three different codes that reflected the use of MKT. For this participant, the items appeared to measure MKT, but the total score across all items was not indicative of her teaching competence.

Table 20.
Verbal Report Excerpt for Retrospective Question 5: Some people would say that a person who correctly answers these PMKT items could be more competent in teaching mathematics than someone who cannot answer these questions. What do you think about that statement?

MATH 207 Participant
P = I am going to say no because I don't know if I necessarily answered them all correctly, but I think that I am just as competent to teach than anyone else is. It is really hard to verbalize exactly you thought. I think if somebody like I know me personally if I sat down by myself and read them in my head and wrote them all out, then I could figure them out and figure out my basic way of teaching them to someone, whereas somebody else could read this and know all the answers. So, I think that everyone has their different way of learning and teaching. So, I think that this wasn't exactly the best way to decide if someone is capable to teach math or not. Some people just are not good test takers.

P = the participant

Emergent Codes. Following the analysis of the a priori codes, emergent codes were obtained from recurring themes in the qualitative data. In addition to the transcribed verbal reports, the principle investigator's memos from the interviews were used to develop the emergent codes. These emergent codes were developed by the principal investigator and discussed with the second coder. These codes existed in two categories: Contextual performance of items and item functioning.

The only code within the contextual performance of items category was "the existence of the student." Across all preservice teachers' responses to the items, the student embedded within the context of the item was mentioned at a much higher rate than the participants in the comparison group. Participants in the comparison group rarely mentioned the student embedded in the context of the item. Instead they focused on the mathematical procedure that could be used to respond to the item. In most cases the student in the comparison group cross checked the response options with the mathematical procedure involved in the item to determine which response was correct. In some cases this process worked (e.g., Item 14), in others it did not (e.g., item 23). Item 23 did not have a clear mathematical procedure that needed to be performed to correctly respond to the item.

Item functioning addressed the following codes: unfamiliar vocabulary, unclear perspective, and ambiguous illustration. Unclear vocabulary addressed issues where the participant was unsure of the vocabulary used in the item. Unclear perspective addressed occurrences in which the participant verbalized uncertainty about what they should focus on within the context of the problem. Ambiguous illustration addressed concerns regarding the interpretation of the illustration (i.e., picture or diagram) embedded in the problem. The participant's interpretation of the item was analyzed and the appropriate code was applied. An analysis of the feedback regarding item functioning (Table 21) was conducted to determine participants' interpretations of the item and the problem experienced in responding to the item. For example, in item 18 the interpretation of the item differed between two participants. The problematic interpretation was associated with the participant's unfamiliarity with the phrase *invert and multiply*. Instances similar to these examples were noted for each item.

Table 21.
Analysis of Item Functioning Based on Cognitive Interview Data

Item Description	18. Which word problem and solution illustrates the invert and multiply procedure.	16. Determine which student is demonstrating dividing parts of a whole.	3. Link representations to mathematical ideas
Interpretations	Identify response that uses this procedure; unable to interpret.	Perform the procedure to identify the correct answer; Determine how the student in the problem uses this procedure.	I have no idea what this picture means, this makes no sense; Figure out if this procedure works for this picture.
Problem Type	Unfamiliar vocabulary	Unclear perspective	Ambiguous illustration

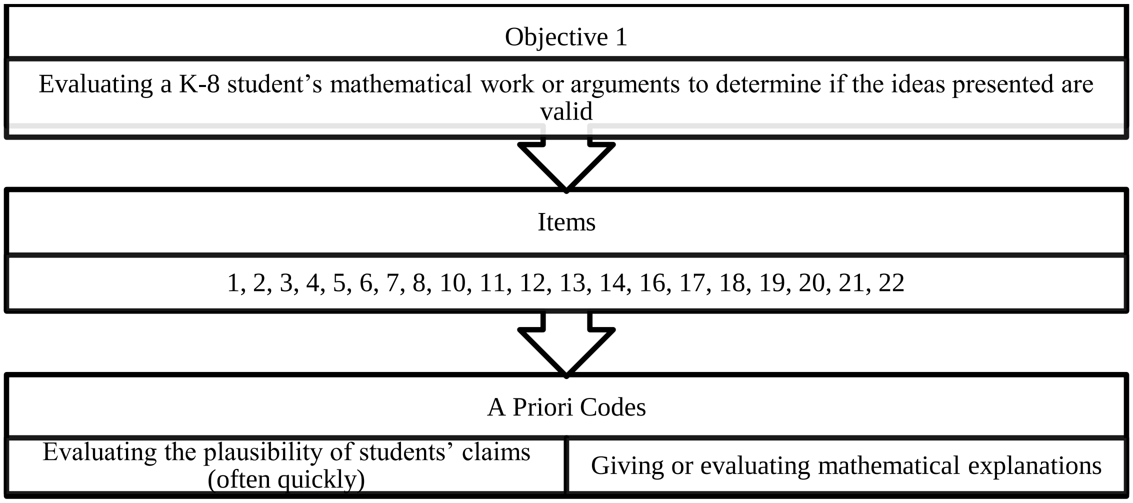
Qualitative Results Summary

Summary of Research Question 3

The a priori codes provided a clear indication that the items and the retrospective questions were eliciting responses related to evaluating the plausibility of students' claims and giving or evaluating mathematical explanations. The use of this type of knowledge was most evident in the upper level preservice teacher courses. While thinking aloud about the items, these preservice teachers discussed their process for evaluating students' claims on many of the items. However, the participants' ability to articulate the knowledge domain that was required for correctly responding to the items was not an indicator of their performance on the test. This was evidenced by the ELED 433 participant whom performed poorly on the items, but was able to articulate the knowledge the items were created to elicit.

Summary of Research Question 4

Analysis of participants' responses to retrospective question 1 provided some indication that the items were addressing a priori codes *evaluating the plausibility of students' claims* and *giving or evaluating mathematical explanations.* These codes were highly related to learning objective 1. It was evident that many of the same items that were mapped to objective one by the content experts were items that were coded using similar a priori codes Figure 4 illustrates this relationship. Taken together, the qualitative data provided indication that the participants believed the items were measuring MKT. More specifically, they cited the evaluative process of MKT as the underlying knowledge for responding to the items. These conceptualizations as verbalized in the transcribed reports were evidence that the items have face validity for this sample of preservice teachers.



Objective 1

Evaluating a K-8 student's mathematical work or arguments to determine if the ideas presented are valid

Items

1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22

A Priori Codes

Evaluating the plausibility of students' claims (often quickly)

Giving or evaluating mathematical explanations

Figure 4.*Relationship between Objective 1, PMKT items, and A Priori Codes*

CHAPTER FIVE

DISCUSSION

The purpose of this study was to develop a scale to advance the assessment of preservice teachers' MKT. To that end, this researcher developed a scale intended to measure the SCK domain of MKT as specified in the learning objectives of a teacher education program. Scores on the items were intended to provide the program with feedback on student achievement on the specified learning objectives. This study was necessary due to a lack of validity evidence for other measures used to assess preservice MKT. In order to validate the use of the PMKT items developed for this study, validity evidence needed to be obtained. This chapter examines the findings of the instrument development and the construct validation results presented in Chapter Four. The quantitative findings are discussed within Benson's stages of construct validity, while the qualitative findings are used to supplement that discussion. Included in this discussion are findings related the study's research questions, implications for future research and practice, limitations, and concluding remarks.

Relationship of the Results to Previous Research and Theory

The development and validation of the PMKT items was intended to provide teacher educators with a program assessment tool. In Chapter 2, the existing research regarding mathematical knowledge for teaching was reviewed. There were two gaps in the literature that were relevant to the development of a preservice MKT measure: defining the construct and developing items for program assessment.

Initial efforts to objectively measure MKT focused on exploring the subdomains of subject matter knowledge and pedagogical content knowledge (Hill et al., 2004). Under these two domains existed several additional theorized knowledge domains (Ball et al., 2008), many of which have not been objectively measured. While the theoretical consideration of these domains were important milestones in MKT literature, there existed several measurement problems. Most noteworthy of the measurement problems was the ill-defined boundaries between the domains of MKT. Indeed, as any assessment practitioner will attest, shoring up the boundaries of a construct is essential prior to assessment.

As suggested in Kane (2007), this research attempted to "tighten the test specifications for the SCK" to reduce the impact of the constructs ill-defined boundaries on measurement (p.185). For the purpose of the current study, the specialized content knowledge domain was operationalized using only learning objectives of the program. Items were only written to measure the portion of the SCK domain that was explicated in the learning objectives. This was done for two reasons. First, it allowed the researchers to hypothesize an essentially unidimensional measurement model for these particular PMKT items. Second, it allowed the teacher education program to clearly align the PMKT items to the objectives of the program. This process helped to clarify the domain of knowledge being measured by the PMKT items.

Previous research on the operationalization of SCK was integral to conceptualizing and making explicit the learning outcomes of the program. Unlike the previous studies, this study focused on the measurement of SCK for program assessment. The PMKT items were used to determine the effect of the teacher education math curriculum on preservice teacher attainment of the program's learning outcomes. Therefore, the validation evidence needed for making inferences about preservice teacher attainment of SCK should be made in the context of program assessment.

Performance of the PMKT items

The first research question addressed the structural component of a strong program of construct validity. The factor structure of the PMKT items were examined to determine whether the unidimensional model of specialized content knowledge would hold. Of the 23 initial items included in the initial confirmatory factor analysis, 11 items were retained. A content analysis of the retained items indicated that they were closely related to a learning objective for evaluating the plausibility of student's mathematical ideas. This objective was created to represent a mathematical task of teaching. The literature on MKT suggests that this task is an operation of specialized content knowledge (Ball et al., 2008). Therefore, the factor structure provides some evidence that the items are measuring a component of specialized content knowledge.

Although an essentially unidimensional scale could be formed from the retained items, the amount of variance in the items that could be explained by the factor was small. This is indicative that something other than the evaluation factor is influencing variations in item responses. This variance in item responses that is unexplained by the evaluation factor is considered error. Identifying ways to reduce the amount of error will improve the precision of the 11 item scale, as well as improve the functioning of the 12 items that were removed from the scale.

The item analysis further addressed the functioning of the PMKT items. Six of the items that were retained for the unidimensional scale demonstrated relatively low difficulty (i.e., *p*-values >.70). This low difficulty is problematic considering that over 50% of the preservice sample is derived from students whom had yet to complete their first course in mathematics education. The discrimination for nine of the eleven retained items was between .20 and .29. This indicates that the relationship between scores on the item and the examinees' total scores was relatively weak. This finding reiterates the findings from the factor analysis. If the items are measuring SCK, then the items should be able to discriminate between participants with varying levels of this knowledge. However, if error exists in the measurement of SCK, then the PMKT items' ability to discriminate amongst participants will be adversely impacted. Consequently, responses to the PMKT items may not provide a precise indication of the participants' specialized content knowledge.

A further investigation of the responses was provided by the distractor analysis. This analysis showed that there were several response options not chosen by the highest and lowest performers on the items. Research on the effectiveness of distractors suggests that no more than three or four are necessary for multiple choice tests (Downing, 1993). Six of the 23 PMKT items had more than four response options. Fifteen of the items had at least one response option that was chosen by less than 5% of the preservice teachers. Having ineffective distractors can negatively impact item discrimination (Haladyna & Downing, 1993). Decreased item discrimination reduces the precision to which the item is able to measure SCK.

Group Differentiation

There were significant differences between preservice teachers in ELED 433 and MATH 107 as well as differences between MATH 207 and MATH 107. These differences were expected. Preservice teachers were expected to exhibit higher levels of SCK as a result of completing higher level courses within the mathematics education curriculum. The lower level courses used in this study were perquisites for the higher level courses. Therefore, the mean scores on the PMKT items were expected to increase from MATH 107 to MATH 207 and from MATH 207 to ELED 433. Though the significant differences were found between MATH 207 and MATH 107 and ELED 433 and MATH 107, there were no significant differences between MATH 207 and ELED 433. This unexpected lack of difference between the latter two courses suggests that the mean scores by preservice teachers in each course similar.

Finding some significant differences among the participants across the courses was encouraging. The magnitude of those differences was relatively small given conventional standards (Cohen, 1992). Effect sizes help to determine the magnitude of those differences. However, the effect sizes do not have an inherent meaning. They must be interpreted within the context of the study. In this study there were two factors that could have contributed to the obtained effect size: the strength of the treatment provided by the course and the precision to which the items can measure the effects of the treatment.

The precision of measurement or lack of random error in your measurement directly impacts the magnitude of the differences that can be observed between groups. When measurement error exists, scores within each of your groups can fluctuate greatly due to a lack of precision, thus obscuring any effect of the treatment group (Thompson, 2006). In this study, the grouping variable was level of math course. Improving the reliability of the PMKT items by addressing item quality is one such way to reduce the errors in measurement that may mask the effects of the preservice mathematics education courses.

A common theme throughout analysis of these items is the identification of error in the measurement of SCK. These errors are likely the cause of the small effect of the mathematics education courses on the participants' SCK. Consequently, the inferences regarding the items' measurement of SCK is adversely impacted. However, there is also another plausible confound in our ability to identify differences in SCK. This confounding variable stems from a lack of control over the treatment provided to the participants. Each mathematics education course had multiple sections. Ensuring the systematic efficacy of each course section in providing learning experience that will assist preservice teachers in meeting their outcomes was not possible. A lack of effectiveness of the course experiences is plausible. However, identifying this lack of effectiveness is predicated on the precision of measure used to assess the knowledge that was expectedly gained as a result of the course experiences.

A major tenet of program assessment is identifying the effectiveness of the curriculum on student learning. In order to identify this effectiveness, the program must be able to identify measures that are sensitive enough to detect the student learning. They must also ensure that their curriculum is providing the experiences that will assist students in meeting their learning outcomes. Once validity evidence for the measure is obtained for the purpose of program assessment, it can then be used to determine the effectiveness of the program. For this reason, the construct validity of the PMKT items was examined prior to investigating group differences.

Just as the learning objectives influenced the item writing, they also should influence the learning experiences provided to students. The mathematics education courses were expected to assist the participants in improving their mathematical knowledge for teaching by focusing on the learning outcomes of the program. However, individual teacher educators are allowed to use the process they deem most efficacious in assisting students in meeting these learning outcomes. This is a fact of program assessment, and should be considered when interpreting the results.

Qualitative Inquiry

Qualitative methods were used in this study to supplement the quantitative approach to examining construct validity. Unlike Benson's stages of construct validity, the qualitative methods herein take into account the participants' voices. The think-aloud interviews allowed the researcher to interpret the participants' experiences when responding to the items. This information added value to the discussion of construct validity by relating examinee thought processes to the construct under investigation.

One finding from the qualitative data suggested that the clarity of the items was an issue for some of the preservice teachers. In some cases, the wording of the item or the illustration embedded in the item appeared to negatively impact the participant's ability to verbalize their thought process for responding to the item. The lack of clarity due to the wording of the item did not occur often. In most occurrences, the lack of clarity was associated with a lack of knowledge about the mathematics required to respond to the item.   Similar to the wording issue, the ambiguity of the illustrations did not occur often. Analysis of the interview transcripts also revealed that some participants thought that some illustrations were ambiguous. It is also plausible that any ambiguity perceived by the participants was related to their lack of knowledge for analyzing illustrations. If so, the qualitative data would suggest that the item was eliciting the knowledge it was developed to elicit (i.e. SCK).

Another finding from the interviews indicated that some preservice teachers used procedural knowledge (i.e. Common Content Knowledge) to identify correct answers to the items. When this occurred and a correct response was obtained, inferences about the measurement of SCK by the PMTK items were undermined. This phenomenon is a plausible cause for some items being equally difficulty across all levels of the mathematics education courses.

When the CCK approach did not work, some participants engaged in verbal thought processes that were indicative of SCK. Use of this knowledge did not guarantee success on the item for these participants. Regardless, the verbal responses elicited by the item provided some indication that SCK was being used. This is the knowledge the items were developed to assess.

The themes that emerged from the retrospective interview data provided some context for the factor analysis, item analysis, and group differentiation results. The factor analysis indicated that an essentially unidimensional scale of SCK was plausible. Unfortunately, the saliency of the factor underlying the scale was weak. However, the themes derived from the interview transcripts suggest that the items were eliciting thoughts about mathematics that were consistent with SCK. This was an encouraging step forward for item development.

The item analysis illustrated patterns of item difficulty that were expected, as well as other that were not expected. Given the mathematics education curriculum, it was expected that the items would become less difficult for participants in the higher level courses. This was the case for most items. However, some items appeared to become more difficult for participants in ELED 433 than for participants in MATH 107. The interview transcripts indicated that preservice conceptualization of MKT varied across and within levels of the math courses. It is likely that phenomenon likely affected the results from item analyses conducted on the PMKT data.

Despite the score variation within and across course levels, one theme was particularly relevant across high scorers on the items. Each of these participants' responses to the retrospective interview questions indicated an understanding of MKT that extends beyond procedural knowledge. This should be expected given the focus of the math curriculum, which urges preservice teachers to formulate a more complex understanding of mathematics. The clearest indication of this complex understanding was exemplified by the participants whom were able to verbalize the need to evaluate the robustness of student procedures to correctly respond to the items.

A concern arose from the retrospective interviews about the relationship between participants' competence as indicated by the PMKT items and their sense of self-efficacy in teaching mathematics.

There were a few interview participants who indicated that the items were measuring more than their procedural mathematics knowledge. Those same participants also indicated that their scores were not indicative of their ability to teach mathematics. This occurrence was noted previous research which suggested preservice teachers can be confident in their teaching ability spite of recognized deficits in their specialized content knowledge (Swars et al., 2007). This issue brings to bear the purposes of program level assessment (i.e. PMKT items) versus the purpose of individual achievement testing (i.e. licensure exam). Unlike licensure exams, the PMKT items were developed to provide the teacher education program with data concerning their preservice teachers' MKT for purpose of program assessment. Therefore, in spite of each preservice teacher's self-efficacy for teaching mathematics, the teacher education program will have objective information about their MKT. This information would assist in improving the program curriculum and thereby help produce more knowledgeable math teachers.

Overall, the psychometric properties of the items suggest needs for revision. Improvement of item discrimination is a major need. Items could be improved by reducing the number of response options. Unnecessary response options can burden the test taker. This burden could affect examiner motivation in low-stakes testing. Also, focusing scale improvement on clearly aligning items with the *evaluation of mathematical claims* learning objective may increase the precision in which SCK is measured by the PMKT items.

Implications for Future Research and Practice

The findings of the current study have many implications for research on the assessment of MKT within teacher education programs. For future research, confirmatory factor analyses need to be conducted on the PMKT items. Data from a new sample of preservice teachers from the same teacher education program should be analyzed to confirm the factor structure championed in this study. Best practices suggest following up exploratory factor analyses with confirmatory factor analysis (Finney & Bandalos, 2010). This approach allows the researcher to test the fit of a hypothesized model of a scale's structure. As items are revised, added, or removed, a confirmatory approach to factor analysis would be a useful tool in assessing the factor structure.

As suggested in Benson (1998) a structural equation modeling approach could also be used to address the relationships between preservice MKT and other constructs such as quantitative reasoning, beliefs about teaching mathematics, and teaching self-efficacy.   This process allows the researcher to test models of the relationships amongst these construct, thus improving our understanding of the MKT in preservice teacher samples. However, before assessing these relationships the PMKT items should be improved to increase the reliability of the measure. Research suggests that the quality of your measures will impact the quality of your statistical model (Deshon, 1998).

The results of this study also provide implications for practice. For teacher education programs, the process used in developing the PMKT items could be useful in developing other measures that assess the learning objectives of their program. As demonstrated in previous research, the use of assessments that are not specifically matched to the learning objectives of the program is not advisable. The development of items that align with their learning objectives and curriculum provide the best case scenario for program assessment. This research outlines a process for doing just that.

For example, the use of distractor analyses provides an opportunity for researchers to explore preservice teachers' misconceptions. Identifying distractors that are commonly chosen by those that score high and those that score low provides an opportunity for the program to address common misconceptions through curriculum changes. The use of results from these analyses to make informed changes to the curriculum is an essential part of program assessment.

Furthermore, the development of the PMKT items negotiated tensions between construct underrepresentation and construct irrelevance. While the make-up of MKT is debated in research, this study narrowed the domain of the construct to focus on components of MKT that are relevant to the learning objectives of the program (i.e. SCK). In doing so, the development of the PMKT items allow the program to focus on measuring the expected learning outcomes of the program. Future research can focus on developing measures of other domains of MKT.

Limitations of the Current Research

As with any research study, there are limitations to the inferences that can be made about the construct under investigation. Previous literature provided a vast and complex representation of the MKT construct domain. To reduce the complexity of MKT as suggested by Kane (2007), items were only created to measure knowledge associated with the subdomain of specialized content knowledge. This purposeful reduction of the construct domain, however, does not negate the existence of other subdomains (e.g. common content knowledge and knowledge of student and content). The relationship between these subdomains can impact the interpretation of scores on the PMKT items. Construct validation requires one to empirically test what the construct is, what it is not, and its relationship with other constructs. While this study helped answer those questions for the PMKT's subdomain of SCK, it did not address those questions relative to the entire construct of MKT.

Though an essentially unidimensional scale of SCK could be configured from the items, the shared variance amongst the items were low. This lack of shared variance weakens the inferences that can be made about the construct being measured by the items. Strengthening the inferences from the use of these items requires a reduction of error variance in the responses to the items. Further item development is needed to reduce this error variance, thereby improving validity and reliability of the PMKT items.

The data in this study were collected from preservice teachers enrolled in different sections of each level of math course. There was no way to guarantee that each course provided the same experience to each preservice teacher. Consequently, the treatment could not be standardized. This study did not identify the extent to which a particular section of a course deviated from providing experiences that promote the attainment of SCK.

The interviewing technique used in this study did not allow the researcher to address level of MKT exemplified by participants' responses. The researcher could only note the existence of thematic representations of SCK based on the transcribed data. The depth of this knowledge could not be determined by the coding process used for the qualitative data in this study. Obtaining qualitative data regarding the depth of this knowledge would have allowed the researcher to make stronger statements regarding MKT at differing levels of the math curriculum. However, addressing the depth of this knowledge required interviewing techniques that were not used in this study.

Conclusion

This investigation was conducted to establish construct validity evidence for inferences made from the use of the newly developed PMKT items. There are several conclusions that can be drawn from the results from this instrument development study. The overall findings suggest that limited empirical support was found for making inferences about MKT based on scores from the PMKT items.

Teacher educators need tools to assess their students' attainment of MKT. Without this tool education programs will lack the ability to determine the effectiveness of their curriculum on developing this knowledge. As previous research has indicated, program can no longer assume that test developed outside of their program will provide results that are useful in program assessment. Establishing measures of MKT that relate to their program's learning objectives is crucial. This study provides a process for doing just that.

This is the first study in the existing literature that examines the development of a measure of MKT that is aligned to the learning objectives of a teacher education program. It is also the first study to capture the voices of preservice teachers when responding to multiple choice items for measuring MKT. Therefore comparisons to other scale development studies for assessing MKT at the preservice is not possible. While the PMKT items are not ideal in their functioning, they do offer foundation on which to build better measures and improve our understanding of preservice MKT.

The process of obtaining construct validity evidence often yields results that allow researchers to not only improve their measures, but also improve their methodology for obtaining data from their measures. The use of group differentiation for obtaining construct validity in this study aroused concerns about the effect of course experiences on MKT. The precision of a measure can easily be under- or over-interpreted when the expected impact of the treatment (i.e. math education courses) is unknown. If learning experiences cannot be made similar within sections of the mathematics education courses, researchers must take this into consideration when interpreting results from an assessment of MKT.

In sum, the PMKT items contribute to the existing literature in MTK measurement by introducing a scale that could be used to assess a teacher education program's effectiveness in assisting its prospective teachers in attaining specialized content knowledge. It is hoped that the process of instrument development outlined in this study will inspire additional research in the development of items for assessing other mathematical knowledge for teaching subdomains that are pertinent to the learning objectives of preservice teacher curriculums.

Appendix A

MATHEMATICAL KNOWLEDGE FOR TEACHING ASSESSMENT

Dear IDLS Student,

Today you will be tested on mathematical knowledge for teaching. This knowledge differs from your typical math knowledge used to solve problems. Though problem solving is required in this assessment, the focus is on your ability to use mathematics as it would be used in a teaching environment. Your responses to these items should provide an indication of your ability to use mathematics in the classroom. The items in the test address the mathematics learning objectives across the program (i.e. Math 107, 108, 207, and ELED 433). The scores on this assessment will be reported in aggregate form (e.g. Students in Math 207 obtained an average score of....). The results will be used to identify areas in the program where we can provide further assistance in helping students reach important learning outcomes. Please provide your best effort. Within days following this assessment interviews will be held to discuss your thoughts on mathematical knowledge for teaching. If you would like to volunteer for this focus group please check the appropriate box at the end of the response form.

Please provide all of your answers on the online answer sheet. Use scrap paper and this hard copy to work through the items. At the end of this questionnaire is a short survey concerning the effort you provided and the importance you've placed on doing well on this assessment. Please take the time to fill this out.

Thank you for participating in your IDLS program assessment. To begin go to the website: www.jmu.edu/assessment/springTests.htm.

1. Mr. Brown was working with his class on subtracting large numbers. Among his students' papers, he noticed that some students displayed their work in the following ways:

+3
843 267
-267 270 ← +30
300 ← +500
800 ← +43
843 576
	843 846 876
	-267 -270 -300
	576
		843 843
		-267 -200
		643
		- 60
		583
		- 7

Method A	Method B	Method C

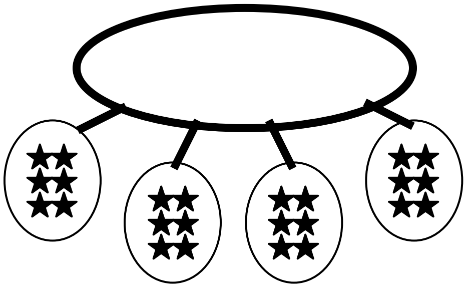
Which of these students is using a method that could be used to subtract any two whole numbers? (Mark ONE answer.)

A. A only **B.** B only **C.** A and B **D.** B and C **E.** A, B, and C

2. Suppose Method A is a correct method. If you were to use this method, what would be your first step in subtracting the numbers 789 – 436?

A. Add 4 **B.** Subtract 4 **C.** Add 3 **D.** Subtract 3

3. Two third-grade students were solving some story problems using counters. Both students started with counters and ended with a display as shown here:



As it turned out, the two students were each doing different problems in the book.

Ricki wrote: $24 \div 6 = 4$

Kyle wrote: $24 \div 4 = 6$

Which of the following story problems was Ricki most likely solving?

A. Rose has 24 Skittles and she wants to give 6 Skittles to each of her friends. How many friends can she give Skittles to?

B. Robby has 24 cupcakes. He wants to share his cupcakes with 4 friends. How many cupcakes will each friend get?

C. Ken has 24 stickers. He wants to give 4 stickers to each of his friends. How many friends can he give stickers to?

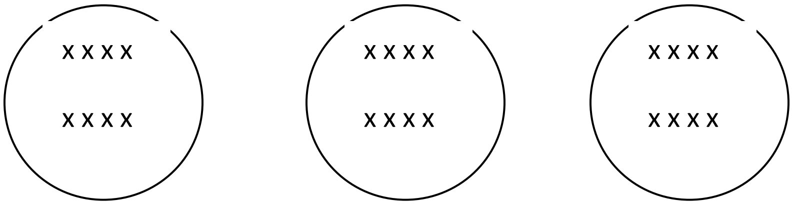
D. Katie has 24 pencils. She wants to share her pencils with 6 friends. How many pencils will each friend get?

4. Students in a third grade classroom were working on the following word problem:

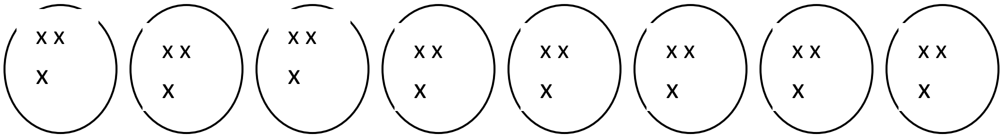
> Annie was baking cookies. She could fit 8 cookies in a pan. If she had 3 pans, how many cookies could she bake?

Which of the following student solutions mathematically models the situation?

A. I drew 3 circles and put 8 dots inside each circle. I then counted and found there were 24. So she can bake 24 cookies.



B. I drew 8 circles and put 3 dots inside each circle. I then counted and found there were 24. So she can bake 24 cookies.



C. Both student solutions are equally a valid way to model this situation.

5. As Mr. Scott was working with his students on subtraction one day, he noticed a few students subtracted in the following way:

$$\begin{array}{r} 12 \\ 8\cancel{2} \\ -\ {}^{4}\cancel{3}7 \\ \hline 45 \end{array}$$

What were these students **most likely** doing? (Choose ONE answer.)

A. The students rounded 37 to 40, then subtracted 40 from 80, and then dealt with the 7 and the 2 in a second step.

B. The students made a mistake with the standard procedure, crossing out the 3 rather than the 8.

C. The students added ten to both the 82 and 37 and then subtracted.

D. The students subtracted 2 from 7, instead of 7 from 2, and then tried to correct for this mistake.

6. A group of second graders are working on the problem 12 – 7=. Jacob solved the problem this way:

```
12  -  7
+3+3

15 – 10 = 5
```

He claims this always works. What do you think? (Choose ONE answer.)

A. Jacob's method only works for certain numbers and he does not realize it.

B. Jacob's method changes the problem to different numbers so it is not correct.

C. Jacob's method only works when one of the numbers is a single digit number.

D. Jacob's method maintains the difference between the numbers and so will work always.

7. Mrs. Smith, a fourth grade teacher, posed the following problem to her students:

The town parade has 25 clowns. Each clown carries 12 balloons. How many balloons were there altogether?

One of Mrs. Smith's students solved the problem this way:

25 x 4 = 100

25 x 12 = 300, there were 300 balloons in all.

What did this student most likely recognize that allowed him to jump from 25 x 4 to 25 x 12?

A. He recognized that he could just use the standard procedure for multiplying two numbers and did that in his head to find 300.

B. He recognized that 25 x 10 = 250 so all he needed was two more 25s or 50 to get 300.

C. He recognized that 3 groups of 4 were 12, so he just tripled the 100 to get 300.

8. A second grade class was working on adding multi-digit numbers using number sense. The following is one student's solution for 58 + 25.

50 + 20 = 70

5 + 5 = 10

10 + 70 = 80

80 + 3 = 83

Where did the student get the 5 + 5 from?

A. She added 8 + 5 and got 13, which is 10 and 3 more. She just showed the 10 using 5 + 5.

B. She split the 8 from the 58 into a 5 and a 3. Then she added this 5 to the 5 from the 25 to get 10.

C. She did not think of the 5 in the 58 as 50, but as 5 and added that 5 to the 5 from the 25 to get 10.

D. It is not clear where she got the 5 + 5 from.

9. Ms. Russell is working with her students on division with fractions and she wants to write a word problem for $4 \div \frac{1}{2}$. Which word problem(s) could be modeled using $4 \div \frac{1}{2}$?

A. Morgan has 4 pizzas and she wants to give half of them to her friend. How much pizza will her friend get?

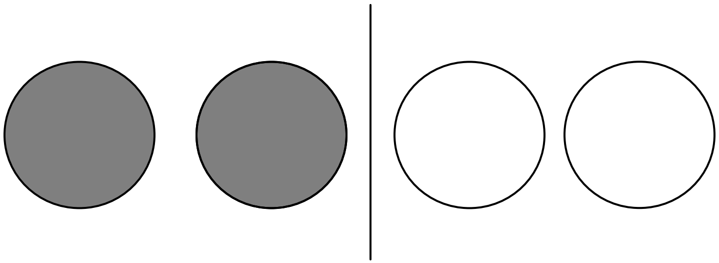
B. Jacob has 4 cups of sugar. He wants to bake cookies, and each batch requires $\frac{1}{2}$ cup of sugar. How many batches of cookies can Jacob make if he uses all of the sugar?

C. Four friends each have half of a muffin. How many muffins would they have if they put them all together?

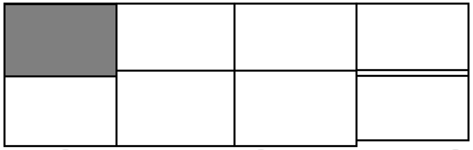
D. All of the above.

10. Some of Ms. Russell's students had drawn pictures to help them solve 4 ÷ $\frac{1}{2}$. Which of the following pictures represents 4 ÷ $\frac{1}{2}$? (Mark ONE answer.)

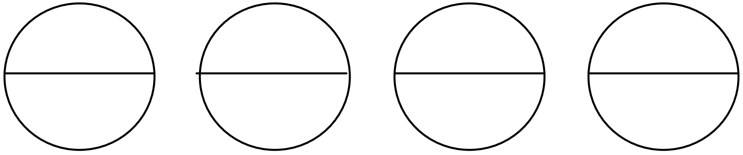
A.



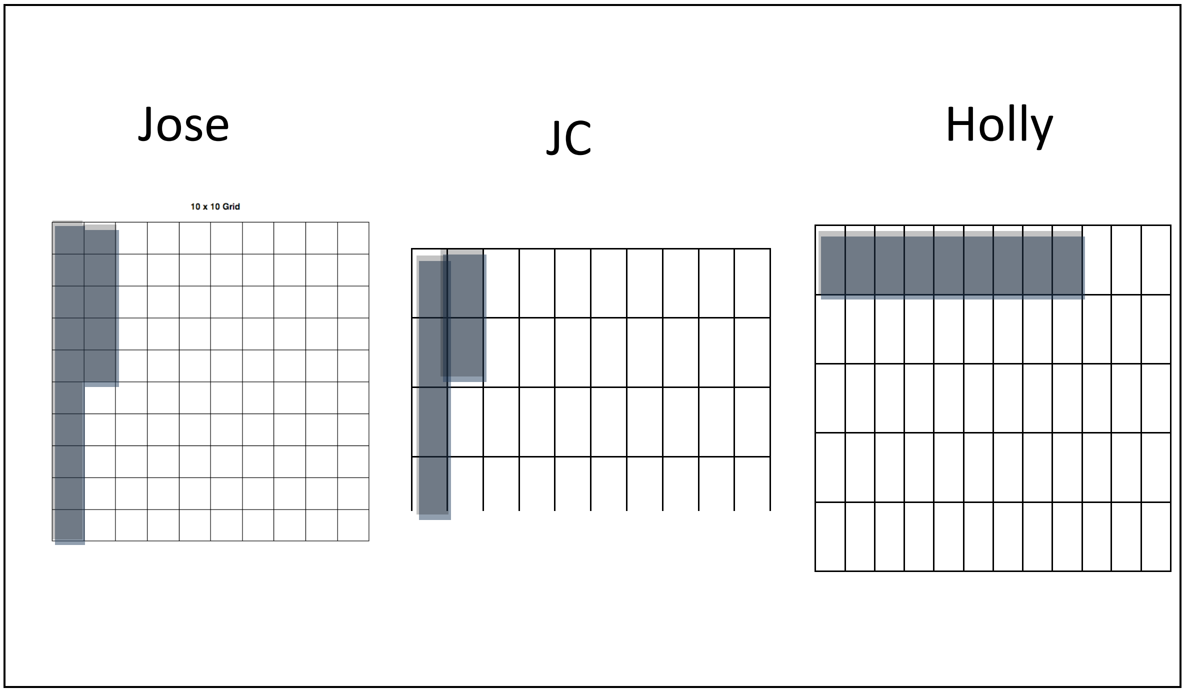
B.



C.



11. Mr. Ramsey was helping his students link fractions, decimals, and percents. Which of the following representations show 15% of the area shaded?



A. Jose only

B. JC only

C. Holly only

D. Jose and JC

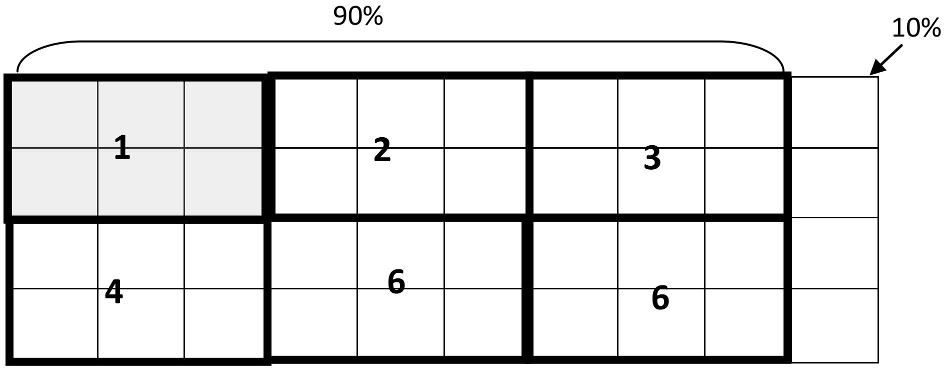
E. Jose and Holly

F. JC and Holly

G. All three.

12. Mr. Ramsey was helping his students link fractions, decimals, and percents. He gave them a 4 x 10 rectangle and asked them to shade 6 squares and determine what percentage of the rectangle was shaded.

Jose drew the following diagram.



Jose reasoned that 90% divided by 6 = 15%, so six squares equaled 15% of the rectangle.

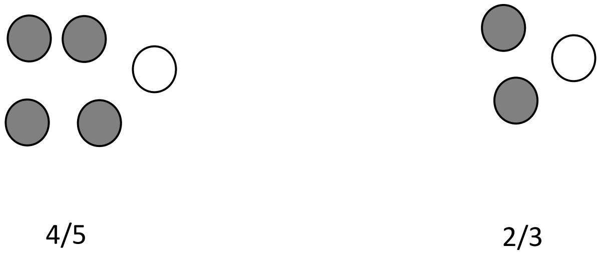
Choose ONE of the following:

A. Jose is incorrect in his thinking because he should have divided 100% by 6, not 90% by 6.

B. Jose is correct in his thinking because the 90% captures 9 rows out of 10 and he has divided 90% of the rectangle into equal parts.

C. Jose is incorrect in his thinking because you cannot find 15% of a 4 x 10 rectangle because there are only 40 squares. You need to divide it into 100 squares before you can find 15%.

13. A student uses the following representation to justify his answer that 4/5 > 2/3.



Since there are more with the 4/5 than with the 2/3, it means that 4/5 is greater than 2/3.

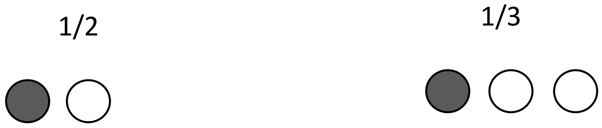
Choose ONE of the following.

A. The student's representation clearly shows that 4/5 > 2/3 because there are more colored counters with 4/5 than with 2/3.

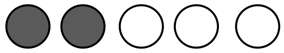
B. The student's representation does not consider that the wholes need to be the same size and so his justification is flawed.

C. The student's representation shows that each fraction is one part away from the whole and that is why 4/5 > 2/3.

14. Suppose a student provides the following argument to show that ½ + 1/3 = 2/5.



Put together or add the ½ and 1/3 to get 2/5.



Choose ONE of the following.

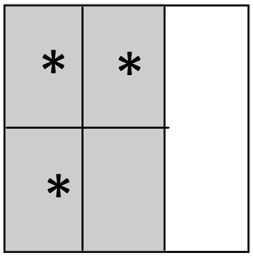
A. The student completed the addition correctly, but his pictures are not correct.

B. The student’s representation shows how to add fractions: add the top numbers and then add the bottom numbers.

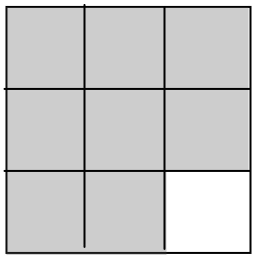
C. The student changes the whole in the situation.

15. Suppose Ms. Chandler wants to use models to help her students think about multiplication of fractions. Which of representation(s) models ¾ x 2/3?

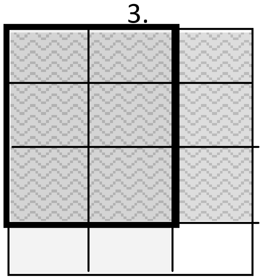
1.



2.



3.



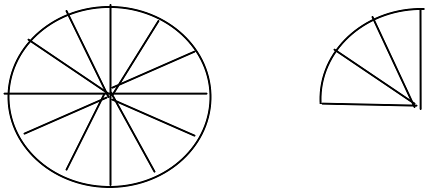
A. 1 only **B.** 2 only **C.** 3 only **D.** 1 and 2 **E.** 1 and 3 **F.** 2 and 3

G. All 3

16. Students solved the following problem:

> Chloe has 1 ¼ hours to finish her three household chores. If she divides her time evenly, how many hours can she give to each chore?

One student solved the problem this way:



There's 15 parts. One third of 15 is 5. So she can give each chore 5/15 of an hour.

Choose ONE of the following:

A. The student is incorrect in his thinking because the whole is 12 parts, not 15 parts.

B. The student is incorrect in his thinking because the 1 ¼ hours should have been represented within one circle. This mistake led him to the incorrect answer.

C. The student is correct in his thinking because he took the 1 ¼ hours and split it evenly among three chores and got 5/15 of an hour.

D. The student's answer (5/15 of an hour) is correct, but his diagram is incorrect.

E. The student's answer and diagram are incorrect. He should have found 3 divided by 5/4, which is 12/5.

17. Mr. Timm's students have been working on comparing fractions. Which explanation(s) is/are correct when comparing 5/7 and 7/9?

1. 7/9 > 5/7 because 7 > 5 and 9 > 7 so 7/9 > 5/7.

2. 7/9 > 5/7 because both are larger than a half but since 7 > 5 then that means 7/9 > 5/7.

3. 7/9 > 5/7 because both are two parts away from a whole, but ninths are smaller than sevenths, so 7/9 has a smaller amount missing from the whole.

A. 1 only

B. 2 only

C. 3 only

D. 1 and 2

E. 1 and 3

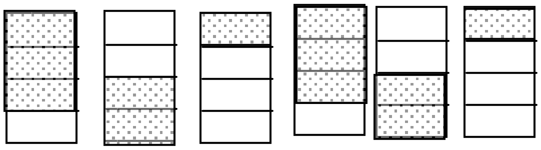
F. 2 and 3

G. All 3

18. Which of the following word problems and solutions more readily illustrates the invert-and-multiply procedure for dividing fractions?

Word Problem 1

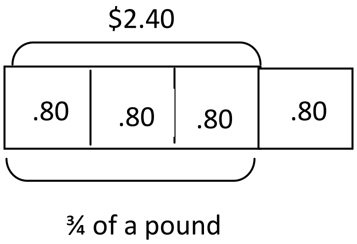
You have bought 6 pints of ice cream for a party. If you serve ¾ of a pint to each guest, how many guests will get ice cream?



I took 6 pints and split them into fourths. I then grouped them into groups of 3. I counted by groups of 3 and had 8 groups. So 8 guests will get ¾ of a pint of ice cream.

Word Problem 2

Olive paid $2.40 for ¾ of a pound of cheese. How much is that per pound?

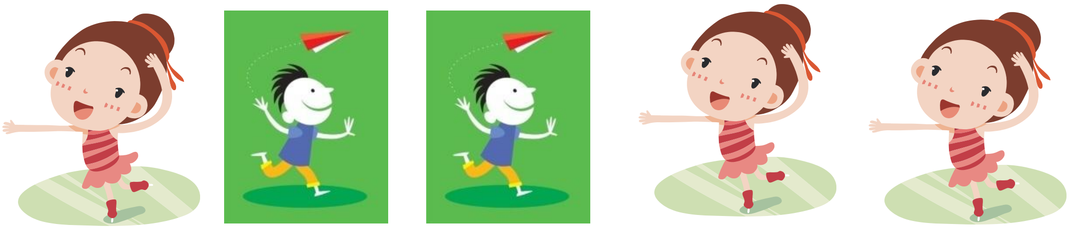


I took the $2.40 and split it into 3 equal parts because I knew it was three quarters. So now I know each quarter is worth $.80. To find a whole pound, I multiplied by 4 to get $3.20.

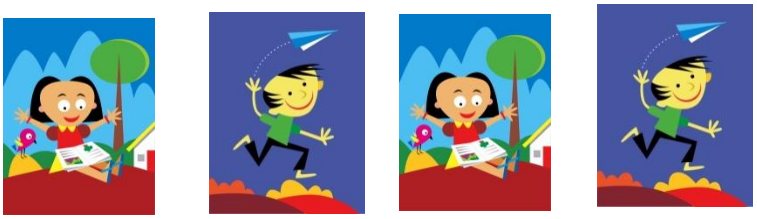
A. 1 only B. 2 only C. Both 1 and 2 D. Neither

19. Mrs. Hardesty is working with her 5th graders on proportional reasoning. She shows them the following picture and asks them which family has more boys.

The Smith Family



The Parker Family



One of her students says that the families have the same number of boys. Mrs. Hardesty has three thoughts come to mind:

1. While the student is correct that both families have 2 boys, the student is not thinking proportionally.

2. The student is correct that both families have 2 boys and is thinking proportionally.

3. The student is incorrect because the Parker family has more boys proportionally.

Which of Mrs. Hardesty's thoughts are mathematically valid? Choose ONE of the following:

A. 1 only

B. 2 only

C. 3 only

D. 1 and 2

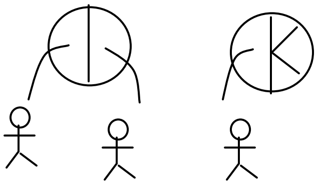
E. 1 and 3

F. 2 and 3

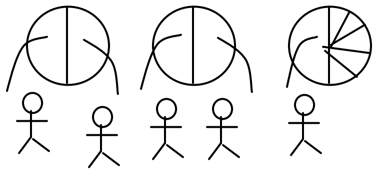
20. A group of students were working on problems in which they were comparing ratios. One of the problems follows:

> Two classes are having pizza parties. Mrs. Adam’s class ordered enough so that every 3 students will have 2 pizzas. Mr. Brown’s class ordered enough so that there would be 3 pizzas for every 5 campers. Did Mrs. Adam’s class or Mr. Brown’s class have more pizza to eat?
>
> One student shared the following representation. What was the student thinking?

Mrs. Adam’s Class



Mr. Brown’s Class



A. Each student in Mrs. Adam’s class will get ½ + 1/3 of a pizza and each student in Mr. Brown’s class will get ½ + 1/5 of a pizza. So students in Mrs. Adam’s class will get more pizza.

B. Each student in Mrs. Adam’s class will get ½ + 1/3 of a pizza and each student in Mr. Brown’s class will get ½ + 1/5 of a pizza. So students in Mr. Brown’s class will get more pizza.

C. Each student in Mrs. Adam’s class will get ½ + 1/6 of a pizza and each student in Mr. Brown’s class will get ½ + 1/10 of a pizza. So students in Mr. Brown’s class will get more pizza.

D. Each student in Mrs. Adam’s class will get ½ + 1/6 of a pizza and each student in Mr. Brown’s class will get ½ + 1/10 of a pizza. So students in Mrs. Adam’s class will get more pizza.

21. Mrs. Olive wanted to see what her students knew about factors. She asked them to write a true statement about the number 12 using the word “factor.” Which of the students’ statements are true? Choose only ONE.

Allie’s statement: 12 is a factor of 6.

Bobbie’s statement: 4 and 3 are factors of 12.

Carrie’s statement: 12 is a factor of 60.

Donnie’s statement: 10 and 2 are factors of 12.

A. Allie

B. Bobbie

C. Carrie

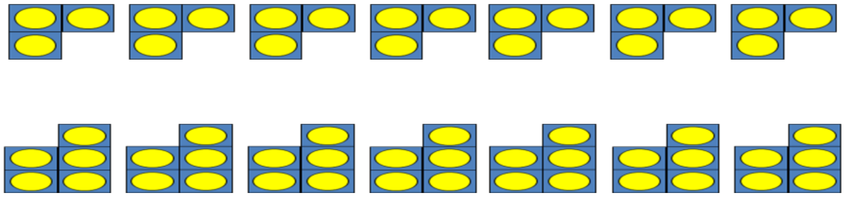
D. Donnie

E. Allie and Carrie

F. Bobbie and Carrie

G. Bobbie and Donnie

22. A student drew this picture to show how he thought about 7 x 8. Which of the following mathematical statements best demonstrates his thinking? Choose ONE.



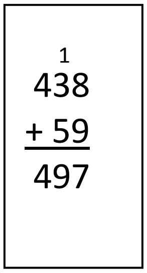
A. 7 x 8 = 7 x 2 x 4 = 14 x 2 x 2

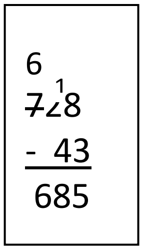
B. 8 x 7 = 7*3 + 5

C. 7 x 8 = 7 x 3 + 7 x 5

D. 7 x 8 = 56

23. Below is the work of Kyle, a second grader, who solved an addition and a subtraction problem.





Does the 1 in each of these problems represent the same amount?

A. Yes, each 1 represents a one because with the standard procedure that is what it is called.
B. Yes, each 1 represents one group of ten because they are in the tens column.
C. No, in the addition problem the one represents a one and in the subtraction problem the one represents one ten.
D. No, in the addition problem the one represents one ten and in the subtraction problem the one represents 10 tens.

Appendix B

Student Opinion Scale

Please think about the test that you just completed. Mark the answer that best represents how you feel about each of the statements below.

A = Strongly Disagree
B = Disagree

C = Neutral

D = Agree

E = Strongly Agree

1. Doing well on this test was important to me.

2. I engaged in good effort throughout this test.

3. I am not curious about how I did on this test relative to others.

4. I am not concerned about the score I receive on this test.

5. This was an important test to me.

6. I gave my best effort on this test.

7. While taking this test, I could have worked harder on it.

8. I would like to know how well I did on this test.

9. I did not give this test my full attention while completing it.

10. While taking this test, I was able to persist to completion of the task.

Appendix C

Think Aloud Prompt

Thank you for signing the consent form and filling out the demographic sheet. We will now begin the think-aloud interview. The purpose of this study is to identify the cognitive processes elicited by the following items concerning learning and teaching mathematics. First, please read **out loud** the instructions proceeding with the items. If anything is unclear, please let me know.

Then, please **read each item stem aloud and verbalize your thinking** as you are contemplating it. Then, for each option following the stem, verbalize your rationale for choosing or not choosing the option. Also, give your overall impression of the item. Feel free **'to speak your mind'** as you are reading the item stem or selecting an option**.** Don't worry about structuring your sentences correctly, but try your best to verbalize every thought that emerges. If something is unclear or confusing, say so. Your responses will be audio-recorded. Please be as honest as you can be and take as much time as you need.

For the research purposes, I will try to keep my interactions with you to the minimum while you are working through the items. Use the scrap piece of paper when necessary. So, if you have any questions, please ask now.

After all items have been answered ask these questions:

- In general, what skills are required to answer the questions you have been presented with?
- Are the skills required to answer these questions being taught in the courses you have taken?
    - If so, in what course or through what actives?
    - If not, do you think they should, and why or why not?
- Do you feel that someone who correctly answers these questions could be more competent in teaching mathematics than someone who cannot answer these questions?

Prompt:

- Explain your process of eliminating the choices you did not choose.

References

Aldridge, J. M., Fraser, B. J., & Huang, I. T.C. (1999).Investigating classroom environments in Taiwan and Australia with multiple research methods. *Journal of Educational Research, 93,* 48-62.

Ambrose, R. (2004). Initiating change in prospective elementary school teachers' orientations to mathematics teaching by building on beliefs. *Journal of Mathematics Teacher Education,7*, 91–119.

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. (1988). Knowledge and reasoning in mathematical pedagogy: Examining what prospective teachers bring to teacher education. *Unpublished doctoral dissertation*. Michigan State University, East Lansing, MI.

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

Ball, D. L. (2000).Bridging practices: Intertwining content and pedagogy in teaching and learning to teach. *Journal of Teacher Education*, *51*(3), 241-247.

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 the teaching and learning of mathematics* (pp. 83-104). Westport, CT: Ablex.

Ball, D. L., Lubienski, S., &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.

Ball, D. L., Sleep, L., Boerst, T. A., & Bass, H. (2009). Combining the development of practice and the practice of development in teacher education.*The Elementary School Journal,109*(5), 458-474.

Ball, D.L., Thames, M.H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389-407.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G.R. Hancock, & R. O. Mueller (Eds.).*The reviewer's guide to quantitative methods in thesocial sciences* (pp. 93-114). New York: Routledge.

Beck, C.T., &Gable, R.K. (2001).Ensuring content validity: An illustration of the process.*Journal of Nursing Measurement*, *9*,201–215.

Behm, S.L, (2008). Preservice elementary teachers' learning with mathematics curriculum materials during preservice teacher education.*Unpublished dissertation.*Virginia Tech.

Begle, E. G. (1972).*Teacher knowledge and student achievement in algebra.*Palo Alto, CA: Stanford University Press.

Begle, E. G. (1979).*Critical variables in mathematics education: Findings from a survey of the empirical literature.* Washington, DC: Mathematical Association of America and National Council of Teachers of Mathematics.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example.*Education Measurement: Issues and Practice, 17,* 10-17.

Borko, H., Eisenhart, M., Brown, C. A., Underhill, R. G., Jones, D., &Agard, 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.

Borko, H., Michalec, P., Timmons, M.,&Siddle, J. (1997). Student teaching portfolios: A tool for promoting reflective practice. *Journal of Teaching Education*, *48*(5), 347-357.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

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

Capraro, R.M., Capraro, M.M., Parker, D., Kulm, G., &Raulerson, T. (2002, January).*Conventional wisdom is wrong: Anyone cannot teach and teachers are not born.* Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, New York.

Center for American Progress. (2010). Measuring what matters: A stronger accountability model for teacher education. Washington, DC: Author.

Chapman, P. (2005). Constructing pedagogical knowledge of problem solving: preservice mathematics teachers.In Chick, *H.* L. & Vincent, S. L. (Eds.), *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education: Vol. 2* (pp. 225-232). Melbourne: International Group for the Psychology of Mathematics Education.

Cizek, G. J. (1991.) Innovation or enervation? Performance assessment in perspective. *Phi Delta Kappan, 72*, 695–699.

Cizek, G. J., & O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement*, 54, 861–872.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Cohen, D. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*(3), 311-329.

Creswell, J.W. (2003). *Research design: Qualitative, quantitative, and mixed-methods approaches*. Thousand Oaks, CA: Sage.

Creswell, J. W., &Plano Clark, V. L. (2007).*Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Toronto: Harcourt Brace Jovanovich College Publishers.

Darling-Hammond, L., & McLaughlin, M. W. (1995).Policies that support professional development in an era of reform. *Phi Delta Kappan, 76*(8), 597-604.

Deshon, R.P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*(4), 421-423.

DeVellis, R.F. (2003). *Scale development: Theory and applications* (*2nd ed*). Thousand Oaks, CA: Sage.

Dewey, J. (1910/1997). *How we think.* Mineola, NY: Dover publications, Inc.

Drennan, J. (2003). Cognitive interviewing: Verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing, 42,* 57-63.

Ercikan, K., Arim, R., Law, D. Domene, J., Gagnon, F., &Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29* (2), 24-35.

Erwin, T. D. (1991). Assessing student learning and development: A guide to the principles, goals, and methods of determining college outcomes. San Francisco: Jossey-Bass.

Fabrigar, L. R., Wegener, R. C.,MacCallum, E. J., &Strahan,D. T.(1999).Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods,*4(3), 272–299.

Flora, D., & Curran, P. (2004).An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods,9*, 466–491.

Frid, S., & Sparrow, L. (2004).Using reflection as pre-service primary teachers develop a mathematics teaching portfolio. In I. Putt, R. Faragher, & M. McLean (Eds.), *Mathematics education for the third millennium: Towards 2010: Vol. 1.*Proceedings of the 27th annual conference of the Mathematics Education Research Group of Australasia(pp. 239-246). Sydney: MERGA.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*, 915-945.

Girod, G. R. (Ed.). (2002). *Connecting teaching and learning: A handbook for teacher educators on teacher work sample methodology*. Washington, DC: AACTE Publications.

Gleason, J. (2010). Reliability of the content knowledge for teaching-mathematics instrument for preservice teachers. *Issues in the Undergraduate Mathematics Preparation of School Teachers: The Journal, 1*, 1-12.

Graeber, A. O., Tirosh, D., & Glover, R. (1989).Preservice teachers' misconceptions in solving verbal problems in multiplication and division. *Journal of Research in Mathematics Education, 20*, 95-102.

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

Grossman, P. L., Wilson, S. M., & Shulman, L. (1989). Teachers of substance: Subject matter knowledge for teaching. In M. C. Reynolds (Ed.).*Knowledge base for the beginning teacher* (pp. 23-36). Oxford: Pergamon Press.

Guba, E. (1987). What Have we Learned about Naturalistic Evaluation? *Evaluation Practice, 8,* 23-43.

Haladyna, T. M., & Downing, S. M. (2004).Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

Hanson, W. E., Creswell, J. W., Clark, V. L. P., Petska, K. S., & Creswell, J. D. (2005). Mixed methods research designs in counseling psychology. *Journal of Counseling Psychology, 52*, 224–235.

Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management, 1,* 19–41.

Hanushek, E. A. (1996). A more complete picture of school resource policies. *Review of Educational Research, 66,* 397–409.

Harbison, R. W., & Hanushek, E. A. (1992). *Educational performance for the poor: Lessons from rural northeast Brazil.* Oxford, England: Oxford University Press.

Hartmann, C. (2004). Using teacher portfolios to enrich the methods course experiences of prospective mathematics teachers. *School Science & Mathematics, 104*(8), 392-407.

Heaton, R. (1992). Who is minding the mathematics content? A case study of a fifth-grade teacher. *The Elementary School Journal, 93*(2), 153-162.

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

Hill, H., Ball, D. L., & Schilling, S. (2008). Unpacking “pedagogical content knowledge”: Conceptualizing and measuring teachers’ topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372-400.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430-511.

Hill, H. C., Dean, C., & 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., & 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.

Ingvarson, L. (1998). Professional development as the pursuit of professional standards: The standards based professional development system. *Teaching and Teacher Education, 14*, 127-140.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33* (7), 14-26.

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

Kagan, D. M. (1992).Implications of research on teacher belief. *Educational Psychologist*, *27*(1), 65-90.

Kahn, J. A., Cooper, D. A., & Bethea, K. A. (2003). The role of mathematics teachers' content knowledge in their teaching: A framework for research applied to a study of student teachers. *Journal of Mathematics Teacher Education, 6*, 223-225.

Kane, M. (2007).Validating Measures of Mathematical Knowledge for Teaching. *Measurement: Interdisciplinary Research & Perspective, 5,* 180-187.

Kaplan, R., & Saccuzzo, D. (1997).*Psychological Testing: Principles, Applications, and Issues.* Pacific Grove, CA: Brooks/Cole Pub. Co.

Knafl, K., Deatrick, J., Gallo, A., Holcombe, G., Bakitas, J., Dixon, J., et al. (2007). The analysis and interpretation of cognitive interviews for instrument development. *Research in Nursing and Health, 30,* 224-234.

Korthagen, F.,& Kessels, J.P (1999). Linking theory and practice: changing the pedagogy of teacher education. *Educational Researcher, 28*(4), 4-17.

Kreber, C., & Cranton, P. A. (2000).Exploring the scholarship of teaching. J*ournal of Higher Education, 71*, 476-496.

Landis, J.R.,& Koch, G.G (1977).The measurement of observer agreement for categorical data.*Biometrics,33*, 159–174.

Lau, A.R., Swerdzewski, P., Jones, A.T., Anderson, R.D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *Journal of General Education, 58*(3), 196-217.

Lawrenz, F.,& Toal, S. (2007). A few tweaks to the toolkit. *Measurement: Interdisciplinary Research and Perspectives, 5*(2-3), 195-198.

Leahey, E. (2007). Convergence and confidentiality? Limits to the implementation of mixed methodology. *Social Science Research, 36,* 149-158.

Leinhardt, G., & Smith, D. (1985). Expertise in mathematics instruction: Subject matter knowledge. *Journal of Educational Psychology, 77*(3), 247-271.

Loucks-Horsely, S., Love, N., Stiles, K., Mundry, S., & Hewson, P. W. (2003).*Designing Professional Development for Teacher of Science and Mathematics, 2nd edition.* Thousand Oaks, CA: Corwin Press, Inc.

Lowery, N. V. (2010). Construction of teacher knowledge in context: Preparing elementary teachers to teach mathematics and science. *School Science and Mathematics*, *102*, 68-83.

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.

Mathews, M. E., & Seamen, W. I. (2007).The effects of different undergraduate mathematics courses on the content knowledge and attitude towards mathematics of preservice elementary teachers. *Issues in the Undergraduate Mathematics Preparation of School Teachers: The Journal, 1*, 1-16.

McConney, A., Schalock, M., & Schalock, H. D.(2001). Focusing improvement and quality assurance: Work samples as authentic performance measures of prospective teachers' effectiveness. *Journal of Personnel Evaluation in Education,11*, 343–363.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Merriam, S. B. (2009). *Qualitative research: A guide to design and implementation*. San Francisco: John Wiley and Sons.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rded.) (pp. 13–103).New York: American Council on Education and Macmillan.

Monk, D.H. (1993). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review*, *13*(2), 125-145.

Morell, L., & Tan, R.J. (2009).Validating for use and interpretation: A mixed methods contribution illustrated. *Journal of Mixed Methods*, *3*(3), 242-264.

Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications to health research. *Qualitative Health Research, 8*(3), 362-376.

Mullens, J. E., Murnane, R. J., & Willett, J. B. (1996). The contribution of training and subject matter knowledge to teaching effectiveness: A multilevel analysis of longitudinal evidence from Belize. *Comparative Education Review, 40,* 139–157.

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

Myers, K. K., &Oetzel, J. G. (2003).Exploring the dimensions of organizational assimilation: Creating and validating a measure. *Communication Quarterly, 51,* 438-457.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future*. New York: Author

National Council of Teachers of Mathematics.(1989).*Curriculum and Evaluation Standards for School Mathematics.* Reston, VA: Author.

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

National Council for Accreditation of Teacher Education(NCATE). (2002). *Professional standards for the accreditation of schools, colleges, and departments of education (*2002ed.). Washington, DC: Author.

National Council for Accreditation of Teacher Education.(2002).*Professional standards for the accreditation of schools, colleges and departments of education*.(2006 ed.). Washington, DC: Author.

National Council for Accreditation of Teacher Education.(2006).*What makes teachers effective?* Washington, DC: Author.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

Nichols, P., & Sugrue, B. (1999).The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice, 18*, 18–29.

Niess, M.L. (2005). Preparing teachers to teach science and mathematics with technology: Developing a technology pedagogical content knowledge, *Teaching and Teacher Education*, *21*(5), 509–523

Putnam, R. T., Heaton, R. M., Prawat, R. S., & Remilliard, J. (1992). Teaching mathematics for understanding: Discussing case studies of four fifth-grade teachers. *Elementary School Journal, 93*, 213-228.

Q.S.R. International (2008).QSR NVivo (Version 8). [Computer Software] QSR International Pty Ltd, Victoria, Australia: Author.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nded.). Newbury Park, CA: Sage Publications.

Pratt, E.O. (2002). Aligning mathematics teacher *w*ork sample content with selected NCTM Standards: Implications for preservice teacher education. *Journal of Personnel Evaluation in Education, 16* (3), 175-190

Quinn, J. R. (1997). Effects of mathematics methods courses on the mathematical attitudes and content knowledge of preservice teachers. *The Journal of Educational Research, 91*(2), 108-113.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 3,* 207-230.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401-412.

Richards, L. (2005). *Handling qualitative data.* Thousand Oaks, CA: Sage.

Richardson, V. (1996).The role of attitudes and beliefs in learning to teach. In J. Sikula (Ed.), *Handbook of research on teacher education* (pp. 102 - 119). New York: Macmillan.

Romberg, T.A. (1995). *Reform in school mathematics and authentic assessment*. New York: SUNY Press.

Rowan, B., Chiang, F., & Miller, R. J. (1997).Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education, 70,* 256–284.

Renaissance Partnership for Improving Teacher Quality [RPITQ], (2002).Homepage. Retrieved from http://www.uni.edu/itq/RTWS/

Russell, J., Goodman, J.T., Anderson, R.D., & Lovin, L. (2010, April).*A psychometric investigation of the learning mathematics for teaching instrument with preservice teachers*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Sale, J. E, Lohfeld, L. H., Brazil, K. (2002) Revisiting the quantitative-qualitative debate: Implications for mixed-methods research, *Quality & Quantity*, *36*, 43-53.

Scott, A. (2005). Pre-service teachers' experiences the influences on their intentions for teaching primary school mathematics. *Mathematics Education Research Journal, 17*(3), 62-90.

Schilling, S. G., & Hill, H. C. (2007).Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement, 5*(2–3), 69–130.

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

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

Simmons, P. E., Emory, A., Carter, T., Coker, T., Finnegan, B., Crockett, D., et al. (1999).Beginning teachers: Beliefs and classroom actions. *Journal of Research in Science Teaching, 36*(8), 930–954.

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

Sleep, L. (2009). *Teaching to the mathematical point: Knowing and using mathematics in teaching.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Smith, J. K. (1983). Quantitative versus qualitative research: An attempt to clarify the issue. *Educational Researcher,12,* 6–13.

Snyder, J., Lippincott, A., & Bower, D.(2000).The inherent tensions in the multiple uses of portfolios in teacher education. *Teacher Education Quarterly, 25,* 45–60.

Spitzer, S.M., Phelps, C.M., Beyers, J.E., Johnson, D.Y., & Sieminski, E.M. (2010).Developing prospective elementary teachers' abilities to identify evidence of student mathematical achievement. *Journal of Math Teacher Education, 14,* 67-87.

Stacey, K., Helme, S., Steinle, V., Baturo, A., Irwin, K., & Bana, J. (2001).Preservice teachers' knowledge of difficulties in decimal numeration. *Journal of Mathematics Teacher Education, 4*(3), 205–225.

Strack, F., & Martin, L.L. (1987). Thinking, judging and communicating: A process account of context effects in attitude surveys. In H.J. Hippler, N. Schwarz & S**.**Sudman (Eds.), *Social Information Processing and Survey Methodology* (pp. 123-148).New York: Springer-Verlag.

Sudman, S., Bradburn, N. M., & Schwarz.N. (1996).*Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

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

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29,* 6–26.

Sundre, D. L.,& Thelk, A. D. (2008, February).*Content alignment and standard setting techniques for general education assessment*. An invited workshop for the General Education Requirements Committee at the University of Miami. Coral Gables, FL.

Suskie, L. (2009). *Assessing Student Learning: A Common Sense Guide (2nd ed*.). San Francisco: Jossey-Bass.

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.

Swars, S. L., Hart, L. C., Smith, S. Z., Smith, M. E., & Tolar, T. (2007). A longitudinal study of elementary preservice teachers' mathematics beliefs and content knowledge. *School Science and Mathematics, 107*(8), 325-335.

Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989). Multiple-choice items: the distractors are also part of the item. *Journal of Educational Measurement, 26,* 161–176.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York, NY: Guilford Press.

Thompson, B.,& Vacha-Haase, T. (2000) Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174-195.

Tourangeau, R. (1987). Attitude measurement: A cognitive perspective. In H.J. Hippler, N. Schwarz & s. Sudman (Eds), *Social Information Processing and Survey Methodology*. New York: Springer-Verlag.

Tourangeau, R., &Rasinski, K.A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299-314.

Vacc, N. N., & Bright, G. W. (1999).Elementary preservice teachers' changing beliefs and instructional use of children's mathematical thinking. *Journal for Research in Mathematics Education, 30*(1), 89-110.

Wade, R.C.,& Yarbrough, D.B. (1996). Portfolios: A tool for reflective thinking in teaching education? *Teaching and Teacher Education, 12,* 63–79.

Wiggins, G. (1991). A response to Cizek. *Phi Delta Kappan, 72,* 700–703.

Williamson, J., & Ranyard, J. (2000).A conversation-based process tracing method for use with naturalistic decisions: An evaluation study. *British Journal of Psychology*, *91*, 203–221.

Willis, G., Royston, P., & Bercini, D. (1991).The use of verbal report methods in the development and testing of questionnaires. *Applied Cognitive Psychology*, *5*, 251–267.

Wise, S. L.,& DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11*, 65-83.

Yu, C., & Muthén, B. O. (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.

Zeichner, K. M. (2000). Ability-based teacher education: Elementary teacher education at Alverno College. In: L. Darling-Hammond, Editor, *Studies of excellence in teacher education: Preparation in the undergraduate years*. Washington, DC:AACTE.

Zeichner, K. M. (2010). Rethinking the connections between campus courses and field experiences in college- and university-based teacher education. *Journal of Teacher Education, 61*(1-2), 89-99.

Zeichner, K. M.,& Liston, D. P. (1996).*Reflective teaching: An introduction*. Mahwah, NJ: Erlbaum.

Zeichner, K. M.,& Wray, S. (2001). The teaching portfolio in US teacher education programs: What we know and what we need to know. *Teaching and Teacher Education, 17*, 613–621.