Exploring the impact of teachers' participation in an assessment-standards alignment study.

EXPLORING THE IMPACT OF TEACHERS' PARTICIPATION IN AN ASSESSMENT-STANDARDS ALIGNMENT STUDY

A Dissertation Presented

By

ANDREA MARTONE

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
Of the requirement for the degree of

DOCTOR OF EDUCATION

May 2007

Education

EXPLORING THE IMPACT OF TEACHERS' PARTICIPATION IN AN ASSESSMENT-STANDARDS ALIGNMENT STUDY

A Dissertation Presented

By

ANDREA MARTONE

Approved as to style and content by:

Linda L. Griffin, Co-Chair

Stephen G. Sireci, Co-Chair

Gloria DiFulvio, Member

Christine B. McCormick, Dean
School of Education

DEDICATION

To Kevin, my loving and supportive husband, I truly could not have accomplished any of this without you. And to our son, Patrick, thank you for sharing me with my work and always having a "huggle" ready for me at the end of the day.

ACKNOWLEDGEMENTS

With deepest gratitude I would like to thank the participants in this study. I appreciated their willingness to participate in all the phases of this study, share their observations and recommendations with me, and support each other as we worked together to examine the issue of alignment. It was a pleasure to work with such dedicated and thoughtful teachers. And my thanks go to Jane Schwerdtfeger, whose support throughout this dissertation project and my work with the adult basic education community has been invaluable.

I also want to thank my advisors, Linda Griffin and Steve Sireci, for their guidance and support throughout my graduate studies and as mentors of my dissertation work. Linda' s comments throughout my dissertation process were instrumental in helping me to more clearly and accurately share my participants' experiences. My dissertation built on my work with Steve in the adult education field and I truly appreciated Steve' s insights and suggestions throughout this study. I was grateful to my third committee member, Gloria DiFulvio, whose questions and suggestions helped me to think about my approach and findings from additional perspectives.

My thanks also go to the professors I worked with prior to my dissertation who helped to shape my interest and approach to research. These professors especially include Patt Dodds, Allan Feldman, Ron Hambleton, Lisa Keller, Judy Placek, and Irv Seidman, I am truly grateful for all of the support and instruction I gained through my graduate studies with each of them.

I would also like to thank my fellow graduate students, Peter Baldwin, Su Baldwin, and Christine Lewis, whose friendship and support made such a difference through this process.

Without the love and support of my parents throughout my life and my education I would not have been able to accomplish what I have. I am also so thankful to the Martone family who welcomed Kevin and Patrick on weekends when I had to really focus and for all of their support throughout this process. And to Rachelle, who played with and nurtured Patrick as I worked, I am truly grateful. Finally, to Kevin, who went above and beyond to support me and our family throughout this process, I am truly blessed to have you as my husband.

ABSTRACT

EXPLORING THE IMPACT OF TEACHERS' PARTICIPATION IN AN ASSESSMENT-STANDARDS ALIGNMENT STUDY

MAY 2007

ANDREA MARTONE, B.S., AMHERST COLLEGE

M.S.T FORDHAM UNIVERSITY

Ed. D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by Professors Linda L. Griffin and Stephen G. Sireci

This study explored the impact of teachers' participation in an assessment-standards alignment study as a way to gain a deeper understanding of an assessment, the underlying standards, and how these components relate to the participants' approach to instruction. Alignment research is one means to demonstrate the connection between assessment, standards, and instruction. If these components work together to deliver a consistent message about the topics about which students taught and assessed, students will have the opportunity to learn and demonstrate their acquired knowledge and skills.

Six participants applied Norman Webb' salignment methodology to understand the degree of alignment between an assessment, the Massachusetts Adult Proficiency Test for Math (MAPT for Math), and state standards, the Massachusetts Adult Basic Education Curriculum Framework for Mathematics and Numeracy (Math ABE standards). Through item-objective matches, alignment was examined in terms of categorical concurrence, depth-of-knowledge consistency, range of knowledge

correspondence, and balance of representation. The study also used observations, discussions, open-response survey questions, and a focus group discussion to understand how the alignment process influenced the participants' view of the assessment, the standards, and their approach to instruction.

Results indicated that the MAPT for Math is well aligned to the Math ABE standards across three out of the four dimensions. Specific recommendations for improvements to the MAPT for Math and Math ABE standards are presented. The study also found that the alignment process influenced the participants' view of the standards, the assessment, and their approach to instruction. Additionally, the study highlighted ways to improve the alignment process to make the results more meaningful for teachers and test developers. This study indicated the value in ensuring an assessment is well aligned to the standards on which it is based.

Findings also showed the value added when teachers are involved in an in-depth examination of an assessment and the standards on which that assessment is based. Teachers are the conduit through which the next generation is guided. Thus it is critical that teachers understand what they are being asked to teach their students and how that can be assessed on a well designed assessment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

There has been a great deal of discourse and debate (i.e., professional and political) regarding the issues and concerns related to testing and standards (Cavanagh, 2004; Darling-Hammond, 2003; James, 2004; Kauffman, Johnson, Kardos, Liu, & Peske, 2002; Linn, 2000; Luna & Turner, 2001; McGehee & Griffith, 2001; Petit, 2002; Popham, 2004). The majority of the debate outlines the negative influence testing has on curriculum implementation. The main concerns surrounding mandated standardized testing include reduced teaching time, a narrowed curriculum and approach to mirror test content, and decreased morale of teachers and students (Smith & Rottenberg, 1991). There has been evidence however to support the view that mandated testing provides a necessary lens to view the educational opportunities presented to students. Without a means to understand what goes on in the classroom and a way to compare how students are performing it is difficult to truly understand if all students are provided with adequate educational opportunities. Well-designed tests have provided important data to learn about student performance and aid in decisions regarding funding, causes of success, and additional options for students and parents (Cizek, 2001).

Even as stakeholders (politicians, educators, parents) debate testing, the psychometric characteristics of the tests have rarely been the basis of their concerns. The primary issue has focused on "opportunity to learn" claims which have weakened the interpretation of assessment results. If a student receives a low score on an assessment, that score should reflect the student' s understanding of the materid taught throughout the year. Unfortunately, students can also receive a low score because they

have not been previously taught the material. Assessments should be a measure of what has been taught throughout the year. Therefore, to refute claims that the tests narrow or disregard the curriculum, research must demonstrate that what is covered on the test supports what occurs or should occur in the classroom, both in terms of the standards and the instruction.

Alignment research is one means to demonstrate the connection between testing, standards, and instruction. If these components work together to deliver a consistent message about the topics students should be taught and assessed, students will have the opportunity to learn and truly demonstrate what they have achieved. The results of alignment research can influence policymakers, assessment developers, and educators to make refinements so these components support each other in what is expected of students. These types of studies have allowed the public to understand how testing does or does not support what is purported to occur in classrooms and what changes may be needed in each of the educational components included in the research.

Standards-Based Assessments

Standards-based assessments are viewed as a way to influence educational reform efforts (Rothstein, 2002). With standards-based assessments the goal is that the assessment both communicates what should be taught and how well it should be accomplished (Herman, 2002). To accomplish this objective in practice it is imperative that teachers are exposed to the content of the assessment and see how the content links to the standards that guide instruction. Alignment research is one means to make these connections explicit. Only if the link between assessment, standards, and instruction is

clear and comprehensive will teachers be confident they are teaching with an understanding of the test rather than blindly teaching to a test.

When the link between assessment, standards, and instruction is in place, in theory, continuous improvement is possible where the results of the assessment can be used to improve instruction and students have the opportunity to demonstrate what they have learned on the assessment (Herman, 2002). If these components are not aligned, if what is taught does not come from the standards or if the assessment does not test what is in the standards, test results will not be able to guide changes that need to be made to improve students' learning.

Standards-Based Assessments' Impact on Teachers

Studies show that teachers analyze what is tested and modify their approach to instruction as needed (Herman, 2002; Lane, Stone, Parke, Hansen, & Cerrillo, 2000). Teachers will match their instructional approach to a test (Stecher, Barron, Chun, & Ross, 2000; Stecher & Borko, 2002) so it is imperative that the test be aligned to what the teachers are expected to teach as described in the standards. As the accountability call gets louder it becomes increasingly important to include specialists from assessment, standards, and instruction in the discussion about how assessments get developed and utilized (Popham, 2004). If test scores are going to be a primary means to show educational gains, then the tests must be based on what the students are expected to learn as stated in the standards. However, beyond just being loosely based, the assessments should show the depth and breadth (in terms of the cognitive complexity and overall standard coverage) of the match to those standards.

While curriculum specialists are more likely to be involved in the assessment development process, teachers are often handed the standards and the assessments and left to discern how these components should be factored into their approach to instruction. The teachers are experts in instructional techniques but may lack the skills to analyze the tests and integrate this understanding with the broad standards. One way to augment teachers' understanding of assessments and standards is through their involvement in alignment research. If teachers were involved in the alignment research, it could be a means to demonstrate the depth and breadth of an assessment as it relates to the standards. In addition, teachers would have an opportunity to interact with the assessment and the standards as each is broken down to its components and connected at the most basic level. Teachers would see how the standards are operationalized in test items as they make connections between the objectives and the items. The process of making these connections might help the teachers to think about what they need to do in their approach to instruction to better meet the scope and depth of the assessment and standards (Blank, 2004).

Approaches to Alignment

The alignment between the test and the standards can be measured through a number of approaches. The Council of Chief State School Officers (CCSSO) recommends three methodologies: Webb (Webb, 1999), Achieve (Rothman, Slattery, Vranek, & Resnick, 2002), and Surveys of Enacted Curriculum (SEC) (Blank, Porter, & Smithson, 2001). Each of these methodologies illustrates the degree of alignment between assessments, standards, and instruction and each is discussed in more detail in the next chapter.

The test development process begins with a development of the test specifications to determine what content and skills the test should measure. Typically the test specifications list the strands and cognitive levels to be measured and then what proportion of the test should be allocated to each dimension. Only if the assessment captures the breadth of the standards can the performance on the assessment generalize to the larger construct to be measured (Rothman, 2003). Alignment methodologies, enacted after the assessment has been developed, help to further define and illustrate the connection to the standards, while also exploring the cognitive demand each item was intended to measure and how this relates to the underlying standards. Alignment research provides validity evidence to demonstrate that the test content is truly representative of state content standards or describes what modifications are needed in either the standards or the assessment to improve the degree of alignment.

For adult basic education (ABE) Mathematics instruction, Massachusetts developed statewide content standards in 2005, the Massachusetts Adult Basic Education Curriculum Framework For Mathematics and Numeracy (hereafter referred to as the Math ABE standards), and a statewide-standardized test that ties each item to an objective within those standards, the Mathematics Massachusetts Adult Proficiency Test (MAPT) (hereafter referred to as the MAPT for Math). At this stage of implementation of both the Math ABE standards and the MAPT for Math, the Webb methodology would best serve the purpose of understanding the degree of alignment between these components as it systematically compares the content in an assessment with the underlying standards across five different criteria. The SEC method, discussed in more detail in Chapter 2, explores how instruction aligns with assessment and/or

standards. This approach will provide better information once the standards and assessment have been in practice for a few years and have had time to impact instruction. The Achieve methodology, also discussed in more detail in Chapter 2, provides a more holistic quantitative and qualitative view of the degree of alignment but again, this approach could provide better information after the Math ABE standards and MAPT for Math have been used for a few years.

At this stage, the Webb methodology will provide concise but specific information that will help the teachers to more thoroughly understand the MAPT for Math and the Math ABE standards, while also providing information to the assessment and standard developers about possible changes that should be implemented. The ABE population, for whom the MAPT was designed, often uses "off-the-shelf" assessments to measure educational gain. The Webb methodology succinctly analyzes four different aspects of alignment of standards and assessments, offers guidelines as to what are acceptable levels, and results in a thorough understanding of strong areas of alignment and areas that should be further addressed.

Purpose of the Study

The purpose of this study is twofold. First, it will explore the degree of alignment between the MAPT for Math and the Math ABE standards. Second, it will examine how teachers' participation in an alignment process influences their views of the standards, the assessment, and their approach to instruction.

Research Questions

1. To what extent is the MAPT for Math aligned to the Math ABE standards?

2. To what extent does teachers' involvement in the alignment process influence their views of the standards, the assessment, and their approach to instruction?

Significance of the Study

This study is significant for four main reasons. First, it builds on the theory of standards-based reform at its initial inception in the adult basic education community in Massachusetts. The theory underlying alignment research is that a consistent message from all aspects of the educational structure will result in systematic, standards-based reform (Smith & O'Day, 1991) where

> an instructional system is to be driven by content standards, which are translated into assessments, curriculum materials, and professional development, which are all, in turn, tightly aligned to the content standards. The hypothesis is that a coherent message of desired content will influence teachers' decisions about what to teach, and teachers' decisions, in turn, will translate into their instructional practice and ultimately into student learning of the desired content (Porter, 2002, p. 5).

Assessments, standards, and instruction are all integral to students progressing through the education system but they have each been determined and enacted separately at multiple levels of the educational structure. While the policy was transmitted through curriculum frameworks documents, different sources created the assessments, and the standards and assessments have been implemented locally in the educational setting through the teachers' individual process for instructing their students. This study resulted in a systematic comparison of the assessment and the standards as a means to compare their content and make judgments about the adequacy of the match and where possible highlight what adjustments might be needed.

Second, this study included teachers as the primary participants and enabled them to become more deeply familiar with the MAPT for Math and the Math ABE standards. Teachers do not typically have extensive exposure to test items and often lack the opportunity to see how those items actually relate to what is written in the standards. Furthermore, as these standards are new to the teachers (developed only six months before the assessment became operational), the teachers do not have a deep understanding of what they are expected to teach to the students or how these standards can guide their instruction (Cohen, 1991). Through the teachers' participation in this alignment study, they systematically analyzed how each item measured an objective within a standard and how the assessment as a whole measured the breadth and depth of the standards.

Third, this study examined the influence the alignment process had on the teachers' thoughts abouttheir approach to instruction given what they learned about the MAPT for Math and the Math ABE standards. As both the standards and the assessments are just being implemented, it was a valuable opportunity to reach out to teachers, involve them in the alignment process, and capture their viewpoint as a means of understanding what next steps might be needed in terms of professional development and enhancing their approach to instruction.

Finally, this study addressed a current gap in the literature in terms of the application of an alignment methodology for a computer-based multi-staged adaptive test in the adult basic education population. Alignment studies are much more common for paper based assessments in the K-12 areas. While computerized-adaptive testing (CAT) is an important way to address the needs of the adult basic education population

given the wide range of skills within a program (Comings & Soricone, 2005), it was critical that important validity issues were addressed to ensure the test is measuring the knowledge and skills as defined in the test specifications (Sireci et al., 2004). The review of the literature shows aligned standardized testing is an area of weakness in the ABE population and CAT testing needs to be more thoroughly evaluated in terms of the content representation at the objectives level. The results of this study demonstrated the content validity of a multi-stage adaptive test tailored to the adult basic education population.

CHAPTER 2

LITERATURE REVIEW

Introduction

A great deal of discourse and debate exist, both professional and political, regarding the issues and concerns related to testing with the majority of the debate outlining the criticisms of standardized testing. The main criticisms of mandated testing in our nation' s schoolsare reduced teaching time, a narrowed curriculum and approach to mirror test content, and decreased morale of teachers and students (Smith & Rottenberg, 1991). There is evidence, however, to support the view that mandated testing provides a necessary lens to view the educational opportunities presented to students. Without a means to understand what goes on in the classroom and a way to compare how students are performing, it is difficult to truly understand if all students are provided with adequate educational opportunities. Well-designed tests provide important data to learn about student performance and aid in decisions regarding funding, causes of success, and additional options for students and parents (Cizek, 2001).

Although politicians, educators and parents debate the merits of standardized testing, the psychometric characteristics of the tests are rarely the basis of concern. Rather, the main criticisms have focused on "opportunity to learn" issues such as testing students on what they have not been taught and narrowing of the curriculum due to mandated testing. Ideally, to address such claims, researchers must demonstrate that what is covered on mandated tests supports what occurs in the classroom, both in terms of the standards and the instruction. Alignment research is one means to demonstrate

the connection between testing, standards, and instruction. If these components work together to deliver a consistent message about what should be taught and assessed, students will have the opportunity to learn and truly demonstrate what they have achieved.

The results of an alignment study can help policymakers, assessment developers, and educators to make refinements so these standards, assessments, and instruction support each other in what is expected of students. Alignment research has allowed the public to understand how testing does or does not support what is purported to occur in classrooms and what changes may be needed in components of the educational system.

Alignment research has resulted in multiple positive outcomes. First, like traditional studies of content validity, alignment studies provide important evidence that can support the validity of test score interpretations (Le Marca, 2001). Second, alignment studies have helped to better understand the number and frequency of content standards currently being assessed and help determine changes that need to be made in future assessments and/or the standards based on content gaps (Ananda, 2003a; Le Marca, 2001; Webb, 1997). In doing so, they address the complaint that large-scale assessments result in a narrowed curriculum. Third, alignment studies have also been used as a legal defense to demonstrate that students are assessed on what they are given an "opportunity to learn" (Webb, 1997) and to compare the assessment approaches among states or districts (Ananda, 2003a). Fourth, alignment research has benefited teachers as they see the connection between classroom instruction and assessments (Webb, 1997) and the results have served as professional development for teachers (Porter & Smithson, 2001). Fifth, alignment studies inform future item writing

activities (Ananda, 2003a), which helps test developers and provides another form of professional development for teachers whenever they are involved in the item writing or item review processes. Sixth, states have used the results of alignment research to inform local planning and decision-making with respect to establishing a baseline to measure future progress (Porter & Smithson, 2001).

Alignment efforts have produced positive outcomes across multiple levels of the educational setting and have allowed all components of the educational field to work toward similar goals to improve student achievement. As Norman Webb, a pioneer of alignment research, stated, "Better aligned goals and measures of attainment of these goals will increase the likelihood that multiple components of any district or state education system are working towards the same ends" (1997, p. 2). Beyond just the alignment of standards and assessments, the instructional content delivered to the students also needs to be in agreement. If this is not the case, if teachers are teaching what they want irrespective of what the standards call for, students could potentially do well in the classroom and then fail on the assessments without understanding where they need additional help (McGehee & Griffith, 2001). Through alignment research, policy makers and educators involved in the educational process can see where they are headed, and will know where they stand relative to an agreed upon goal.

The purpose of this literature review is to describe why an understanding of alignment is an important characteristic of a testing process and how undertaking alignment research can be beneficial both to the participants and the consumers of the results. The review is structured around four areas of discussion. First, an overview of how alignment is defined in the educational measurement literature is presented. This

overview includes formal definitions of alignment and describes how alignment builds on earlier notions of content validity. In the second section, three of the most widely used methods of alignment research are described. While these methods share some common components, a closer look at each approach highlights the relative differences between the methodologies. Specific applications of each methodology are also presented in this section. The third section discusses how alignment research can support teachers and serve as a form of professional development. This section extends the basic alignment research to show how the process itself, more than just the results, can help teachers to see how assessments can connect to what happens in the classroom in a meaningful way. The final section discusses alignment and professional development issues specific to the adult basic education field, which is the population for this alignment study.

Overview of Alignment

Alignment means many things in the educational world. A Webster' s dictionary definition states that to align is "to bring into a straight-line; to bring parts or components into proper coordination; to bring into agreement, close cooperation" (as cited in Le Marca, Redfield, Winter, & Despriet, 2000, p.1). In a classroom setting, instructional alignment refers to agreement between a teacher' s objectives, activities, and assessments so they are mutually supportive (Tyler, 1949). On a school wide level, curricular alignment refers to the degree to which the curriculum across the grades builds and supports what is learned in earlier grades (Tyler, 1949). Alignment, as described in this review, took curricular alignment a step further and looked at "the degree to which expectations and assessments are in agreement and serve in conjunction

with one another to guide the system toward students learning what they are expected to know and do" (Webb, 1997, p. 4). LeMarca et al. (2000) presented a more comprehensive definition of alignment:

> Alignment is defined here as the degree to which assessments yield results that provide accurate information about student performance regarding academic content standards at the desired level of detail, to meet the purposes of the assessment system. To satisfy this definition, the assessment must adequately cover the content standards with the appropriate depth, reflect the emphasis of the content standards, provide scores that cover the range of performance standards, allow all students an opportunity to demonstrate their proficiency, and be reported in a manner that clearly conveys student proficiency as it relates to the content standards (p. 24).

In a perfect world, what a student is tested on should be derived from what is expected of the student as detailed in the school or district curriculum frameworks, as well as what is taught to the student by his/her teachers. While not everything that is listed in the standards or taught to the student can or should be assessed, alignment research has illuminated how much and to what degree the curriculum framework coverage or instructional content has been assessed. An understanding of alignment dimensions is sometimes used at the outset to create curriculum frameworks and assessments that are aligned from their inception (Rothman, 2003). The results of alignment research have been used in conjunction with the priorities determined by educational stakeholders to meaningfully inform future educational decisions.

The theory underlying alignment research is that a consistent message from all aspects of the educational structure will result in systematic, standards-based reform (Smith & O'Day, 1991) where

> an instructional system is to be driven by content standards, which are translated into assessments, curriculum materials, and professional development, which are all, in turn, tightly aligned to the content standards. The hypothesis is that a coherent message of desired content will influence teachers' decisions about

> what to teach, and teachers' decisions, in turn, will translate into their instructional practice and ultimately into student learning of the desired content (Porter, 2002, p. 5).

Assessments, standards, and instruction are all integral to the student achievement, but they have each been determined and enacted at multiple levels of the educational structure. Curriculum frameworks represent policy documents, but sources outside the policymakers created the assessments, and the curriculum and assessments are implemented locally in the educational setting. Alignment studies allow researchers to systematically study the different components of the educational structure as a means to compare their content and make judgments about the adequacy of the match.

Webb noted that the Education Goals 2000 Act supported the development of a consistent message about student learning among the policy, assessment, and instruction perspectives. As he put it, this act " indicated alignment of curriculum, instruction, professional development, and assessments as a key performance indicator for states, districts, and schools striving to meet challenging standards" (Webb, 1997, p. 1). Alignment research has examined the relationship between these educational components, assessment, standards, and instruction. Webb (1999) stated,

> Alignment is defined as the degree to which standards and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between standards and assessments and not an attribute of any one of these two system components (p. 2).

It is not sufficient to understand the benefits of any educational component in isolation so alignment research has focused on how these components work together to send a consistent message about student achievement. Research in this area has examined the multiple dimensions that work together to illustrate the degree of match between

standards and assessments. Additionally, the No Child Left Behind Act (NCLB) requires that a state' s academic achievement standards be aligned with the state' s academic content standards. If the alignment between academic achievement and content standards is low, a state is likely to have trouble meeting the requirements of NCLB. Alignment research culminates in a report about the relationships of the components that can be used for future decision-making rather than just a simple yes or no response (Rothman et al., 2002). The results of alignment research provide a measure of how well assessments cover the underlying content standards and the degree to which assessment and content standards match classroom instruction. Once the degree of alignment is understood, subsequent changes in any of the educational components can be made to improve the standards-assessment-instruction cycle.

In summary, alignment studies provide data that can be combined with the priorities of educational stakeholders to guide changes in assessments, curriculum, and/or instruction. By focusing on the match between test content and what is intended to be taught, alignment research shares some common goals and methodology with traditional methods for studying content validity. The next section discusses some similarities between contemporary evaluations of alignment and traditional studies of content validity.

Alignment as a Form of Content Validity

Generally defined, content validity refers to the degree to which a test appropriately represents the content domain it is intended to measure. When a test is judged to have high content validity, its content is considered to be congruent with the testing purpose and with prevailing notions of the subject matter tested. Thus, content

validity does not specify particular aspects of the educational process such as curriculum frameworks or instruction. Rather, it is more general and refers to tests both within and outside educational systems (e.g., licensure and certification tests).

There are at least four aspects to content validity—domain definition, domain representation, domain relevance, and appropriateness of the test construction procedures (Sireci, 1998a, 1998b). Domain definition refers to the process used to operationally define the content domain tested. In the case of K-12 achievement testing, the domain is typically derived from state-established curriculum frameworks. Domain representation refers to the degree to which a test represents and adequately measures all facets of the intended content domain. To evaluate domain representation, inspection of all the items and tasks on a test must be undertaken. Studies of domain representation typically use subject matter experts (e.g., teachers) to scrutinize test items and judge the degree to which they are congruent with the test specifications (Sireci, 1998b). Domain relevance addresses the extent to which each item on a test is relevant to the domain tested. An item may be considered to measure an important aspect of a content domain and so it would receive high ratings with respect to domain representation. However, if it were only tangentially related to the domain, it would receive low ratings with respect to relevance. Appropriateness of test development procedures refers to all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material. The content validity of a test can be supported if there are strong quality control procedures in place during test development, and if there is a strong rationale for the specific item formats used on the test.

Traditional studies of content validity typically use subject matter experts (SMEs) to rate test items with respect to their congruence to the test specifications or their relevance to the intended domain. Hence, traditional content validity studies and contemporary alignment studies are similar in that they both gather data from SMEs, and structure the data collection procedure in a way that independently evaluates specific aspects of content domain representation.

Sireci, Robin, Meara, Rogers, and Swaminathan (2000) provided an example of a traditional content validity approach to alignment using the Grade 8 1996 NAEP Science Assessment. A primary goal of their study was to evaluate the congruence between the NAEP Science Framework and the NAEP Science Assessment. Ten carefully selected SMEs reviewed a sample of NAEP Science items and were asked to assign each item to (a) one of the three content areas ("fields of science"), (b) one of the three cognitive levels ("ways of knowing and doing science"), and (c) one of the four "themes of science" listed in the NAEP test specifications (framework). Each item was given an item congruence index rating based on the number of raters who agreed with the original classification. For example, if an item was intended to measure Earth Science and 8 out of 10 SMEs rated it as Earth Science, it had an item-content area congruence rating of 0.8. An index of 0.7 and greater was used to judge an item as adequately congruent with its content area, cognitive level, or theme.

While the traditional content validity approach involves rating or matching items to more global levels within test specifications (such as "domains," "strands," or "content areas"), contemporary alignment research uses the same expert rating approach, but delves deeper to examine the match between items and the objectives or

benchmarks within a standard. For example, a state' s curriculum framework may have the standard Grade 4 Number Sense (4N). It is at the level that many test specification tables are written. But within the standard 4N there are multiple objectives. For example 4N-1.1 might be "Read, write, order and compare numbers up to 1,000,000". In this example, the objective provides the detail of what skill the item associated with it should measure. Alignment research often matches items to these detailed objectives and then reports findings summarized by standard. In some cases alignment research has also considered what was actually taught to the students. In this way, alignment research can offer a deeper view of the educational process, and can be thought of as an extension of a content validity evaluation. As is discussed later, however, traditional content validity studies may have some advantages for evaluating the congruence of a particular test form to its test specifications.

Valid assessment requires significant overlap between the assessment and the desired standards to ensure decisions made based on test results are defensible. Alignment research is related to validity, but there is an important distinction that Webb (1997) highlighted: "Validity refers to the appropriateness of inferences made from information produced by an assessment (Cronbach, 1971). Alignment refers to how well all policy elements in a system work together to guide instruction and, ultimately, student learning" (p. 4). Alignment research has been most closely associated with content validity as a means to provide for a common understanding of what students should learn as a guide for instruction and to ensure equity for all students (Bhola, Impara, & Buckendahl, 2003; Webb, 1997). While alignment research examines how well all the aspects of the educational system work together to impact student learning,

validity research focuses on the appropriateness of the interpretations made from the results of the assessment. Thus, alignment research is an example of a validity study needed to support the test score interpretations.

Building on content validity studies, alignment research has helped various state departments of education to systematically compare what has been listed in the standards to what has been tested. In Webb' s(1997) work he found that

> most states' frameworks and assessments were judged to be aligned if goals and learning objectives were considered in the design or selection of the assessment instruments. Most states lacked a formal and systematic process for determining the alignment among standards, frameworks, and assessments (p. 8).

Alignment research has addressed the states' deficiency by systematically comparing the different pieces of the educational process. If educational components are not well aligned, the system will not send a consistent message about what is prioritized in the educational process (Webb, 1999). Thus, alignment research addresses the concerns that the curriculum has been dumbed down (Linn, 2000), that students have not received a fair chance to learn what they were tested on (Winfield, 1993), and that states have not addressed the need to improve instructional quality (Rothman et al., 2002). Alignment research has accomplished this through an analysis of the content of the assessments and to what degree that has matched the goals set forth for the students through the standards and in some cases the instruction.

Approaches to Alignment Research

Alignment research has served multiple purposes. Research on alignment has been used to identify areas of vulnerability based on content gaps, restructure assessments, compare standards and assessments to other states, districts, or localities, inform future assessment item development, and show content validity based on an

objective evaluation (Ananda, 2003a). There has been an expectation that tests should be fair and this has related not just to sensitivity toward the test takers but to the test takers' expectations that thetest content will overlap with what has been taught in the classroom (Crocker, 2003). The degree of overlap has been a crucial aspect of content validity with the alignment of test content and expectations as the only part of the validation process that can occur prior to test delivery and reporting results (Crocker, 2003). As such, it has been important to thoroughly complete the process and accurately report the findings. Alignment research has relied on the development and application of objective methods to determine that the score a student receives on an assessment has been based on performance relative to skills that represent expectations for that domain (Le Marca, 2001).

Some alignment studies have focused on the content of the standards compared to the assessments and others have included the content of instruction as an additional variable. The following section will elaborate on the three most common methods for alignment research – Webb, Achieve, and Surveys of Enacted Curriculum (CCSSO, 2005). An application of each of these methodologies is also presented to illustrate their processes and findings. Throughout this section points of comparison among the three approaches are highlighted.

Webb Methodology

Norman Webb developed a comprehensive and complex methodology to investigate the degree of alignment between assessments and standards. His methodology explores five different dimensions to understand the degree of alignment: content focus, articulation across grades and ages, equity and fairness, pedagogical

implications, and system applicability (Webb, 1997). Each dimension is described below. In this methodology, standards are the broad content domains within a subject and the skills within this domain are referred to as objectives. Understanding these definitional terms is critical to seeing how the alignment process has been applied, because these terms and levels of analyses differ across the different alignment methodologies.

Content focus. Webb' s content focus dimension comprisessix subcategories for analysis: categorical concurrence, depth of knowledge, range of knowledge, balance of representation, structure of knowledge, and dispositional consonance. Each of these subcategories explores the relationship between the assessment and the standards in a different way. Together they contribute to a more thorough understanding of the degree of alignment between assessments and standards than traditional content validity research has provided.

Categorical concurrence compares the similarity of the expectations for student learning, as expressed through the content categories in the standards, to students' assessments. This subcategory of alignment research is most similar to traditional content validity studies and has been a minimum requirement in most alignment methodologies. Like the test blueprint comparison in traditional content validity research, the categorical concurrence variable also looks at the broad content areas (or strands), such as Number Sense and Geometry. To have alignment relative to this variable, an assessment must have had at least six items measuring a standard, defined as the broad content domains. Using this approach, if there are four standards, an assessment needs at least 24 items with six items per standard to determine there was

alignment relative to categorical concurrence. Unlike a traditional content validity study, however, where a test item is matched to its standard by SME consensus (e.g., 70% of SMEs match an item to its intended standard[1]), Webb' s criterion is simply that, across the SMEs, an average of at least 6 items is matched to the standard and there is no requirement for review agreement[2]. That is, a standard could theoretically be considered adequately represented, even if the six items matched to it were specified to measure a different standard in the test blueprint. While Webb uses the criteria of 6 items per standard, the traditional content validity approach compares the actual item representation to the proportions specified in the test specifications for the assessment. Without this comparison, the criterion of 6 items seems to be a minimum requirement at best.

Depth-of-knowledge consistency compares the level of cognitive complexity or type of thinking required as expressed in the specific objectives within each standard to the cognitive complexity in each item that is matched to that objective. Webb initially defined the cognitive areas as recall, skill/concept, strategic thinking, and extended thinking, but these may be modified for a particular study (Webb, 1999). The main criterion here is that what is tested should be at the same cognitive level or above as what is expected to be taught. To have alignment relative to this criterion, at least 50% of the items matched to an objective must be at or above the cognitive level of that objective. Fifty percent is based on the assumption that most cutoff points require students to answer more than half the items to pass but some interpretation is allowed

---

[1] Popham (1992) and Sireci (1998b) suggested the use of 7 out of 10 SMEs correctly matching an item to its intended standard as a criterion for a congruent item-test specification match.
[2] Webb et al. (2006) recently studied the potential impact of enforcing a degree of agreement among participants.

with this point. The main concern in this aspect of alignment is that assessment items should not be targeting skills that are below those required by the objectives.

Range of knowledge correspondence analyzes the breadth of the standards as compared to the breadth of an assessment. This dimension looks at the number of objectives within a standard measured by at least one assessment item. To have sufficient alignment relative to range of knowledge, at least 50% of the objectives within a standard need to be measured by at least one assessment item. This assumes that students should be tested on at least half of the domain of knowledge. This part of the alignment process also assumes all of the objectives have equal weighting and all of the objectives accurately cover the skills needed to complete that standard. The level of complexity within a state' s standards influences this aspect of alignment as more complexly written objectives might only be partially assessed but would still be considered a match from the perspective of this dimension.

Balance of representation focuses on the degree to which items are evenly distributed across objectives within a standard to represent the breadth and depth of the standards. Given a limited time for assessment, this dimension highlights what aspects of the standards are prioritized. Balance of representation focuses on the objectives assessed by the items and then looks at the proportion of objectives measured compared to the number of items. The goal is to measure every objective assessed with at least two items. Specifically the calculation for the balance index is:

$1-(\sum_{k=1}^{O}|1/(O)-I_k/(H)|)/2$, where O=Total number of objectives hit for the standard; I_k = Number of items corresponding to objective k; and H = Total number of items hit for the standard (a hit is any item-objective match)(Roach, Elliott, & Webb, 2005). If the

proportion approaches zero that signifies many items are assessed by only a small number of objectives. If it approaches one it signifies that the items are evenly distributed across all objectives. Ideally, over time, assessments should shift in the balance of representation to cover the entire standards[3]. Evaluating balance of representation across grades can also demonstrate shifts in priorities as the content develops.

These first four areas of Webb' s methodology– categorical concurrence, depth of knowledge, range of knowledge, and balance of representation – are most often used by other alignment researchers as the basis for their alignment methodologies. These four dimensions serve as the most direct way to view the degree of match between an assessment and the standards. The last two aspects of the content focus dimension – structure of knowledge and dispositional consonance – have not been applied in a research study as best as can be determined, but they illustrate the complexity of the alignment process.

Structure of knowledge analyzes to what degree the assessment items target the broader goals of instruction. For example, if the goal is for students to have an integrated understanding of a concept, this variable examines to what degree the assessment is only targeting isolated skills. Webb emphasized that this might best be analyzed in the context of the broader assessment system where it is possible to include both formative and other forms of summative assessments. No researcher to date has integrated this variable into an alignment methodology. Dispositional consonance is another view of structure of knowledge in that it assesses the degree to which the

---

[3] Thus, evaluating the specific standards covered over time is necessary to ensure important standards are not being neglected.

assessments support the broader stakeholder beliefs about education. For example, in the standards it may state that it is important that students be able to critique their own work. This skill is easier to assess in non-standardized settings and highlights the need for alignment studies to include the broader assessment policies of an educational setting. This aspect of content focus would further address concerns about "narrowing of the curriculum."

Articulation across grades and ages. Webb' s method also discusses the need to analyze the change in content across grades and ages as this highlights the content and cognitive complexity in standards. Webb believed that assessments should be developed with an understanding of how students change through the years and how this can be assessed at different stages of development. Cognitive soundness is one aspect of articulation across grades and looks at how the cognitive complexity increases as students move through levels of understanding connecting new ideas to existing ideas. Cumulative growth in content knowledge during schooling is another aspect of articulation across grades and relates to the idea that students start with basic ideas and build on those through schooling. While theoretically these are important pieces of the alignment puzzle, these topics have not been included in alignment research to date although they are important issues that are included in approaches to vertical alignment.

Equity and fairness. Webb highlights issues of equity and fairness in his general approach to alignment. Equity and fairness is a means to ensure high standards are set and every student is given the opportunity to demonstrate understanding. Webb discussed how the form of the assessment could impact this aspect of alignment as some students might respond better to more open-ended assessment tasks. Also, the diverse

backgrounds of students need to be considered in the design of assessments. Finally, this area addresses the concern that students may be at a disadvantage based on the structure of the standards or the developmental level of the student and therefore their achievement level is not related to instruction or student effort. While these aspects of standardized assessments are at the core of the debate about standardized testing, they have not been formally integrated into applied alignment methodologies.

Pedagogical implication. The area of pedagogical implications focuses on the teacher' s interpretations of the curriculumframework expectations and the assessments and how their instruction fits within the context. At times teachers may think they are addressing the standards but in reality they are only superficially meeting the broader expectations (Cohen, 1990). One aspect of the pedagogical implications was teachers' engagement of students and their use of effective classroom practices to send a consistent message about what should be taught and assessed. For example, if the teachers emphasize group work but then this is not assessed it could send a conflicting message. This aspect of alignment again supports the need to look at the broader context of assessments to ensure that the curriculum is not narrowed. Another aspect of pedagogical implications is teachers' instruction regarding technology, materials, and tools. Teachers need to understand what students are expected to do with these materials and this needs to be included in the assessments. In this way again the students will receive a consistent message about what is emphasized.

System applicability. Through system applicability Webb discusses the need to examine how well what is going on in the classroom relates to real world needs. Webb highlights the need for people inside and outside of the classroom (teachers,

administrators, parents, policy makers) to be on the same page with regard to what is valued and focused on in the educational process. This can be fostered through an open assessment process however it is not formally studied in the alignment process.

Methodology application. Webb (1999) applied his methodology in a study of mathematics and science assessments and standards in four states. This literature review focuses on the mathematics alignment process and the results reported from this study. The purpose of Webb' s study was to better understand how his alignment methodology functioned, to examine in greater detail the different alignment variables, and to understand ways to improve the alignment process. In this study, six participants compared the match between assessment items and standards/objectives in mathematics. The results of this matching were used to judge the degree of alignment based on four of Webb' s criteria: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation.

The review process involved multiple decision points by the participants. Applying this process across multiple states, the participants noted differences among the standards in terms of content covered, level of detail for the standards, and the overall organization of the standards, which impacted the comparability of the states. Despite these differences, the first step was a review of each state' s standards in order to match each objective to a depth-of-knowledge level representative of the highest level of knowledge needed to achieve that objective. Systematically matching every objective with an associated depth of knowledge allowed for objective-item matches beyond just the basic content validity approach. The participants reached an agreement about the depth-of-knowledge of the objectives based on a group discussion. These

decisions were used as a baseline comparison to the assessment items to determine if the items were at or above the cognitive level in the objective.

The items within an assessment were then matched to the objectives within the standards and coded based on the depth-of-knowledge required by that item. Any match was called a "hit", however, one item could be matched to more than one objective. This increased the content and range alignment criteria areas but proved to be an area of confusion for the participants. The participants also noted when items appeared to not match any objective and a generic objective was created to match to these items. The results of this study were aggregated to report by standard. The mean and standard deviation for each criterion were computed for each participant.

The results showed varied levels of alignment across grade levels and states. The strongest area of alignment was the categorical concurrence criterion, the criterion most similar to a traditional content validity study. Three out of the four states fulfilled this criterion with at least six items measuring a standard but in each state one-fourth or more of the standards were measured by less than six items. The balance-of-representation criterion was satisfied as standards that were assessed had items that were evenly distributed among the objectives.

The weakest aspects of the alignment methodology were the depth-of-knowledge consistency and range-of-knowledge correspondence criteria. The results demonstrated that assessment items generally targeted a lower level of knowledge and did not sufficiently cover the range of knowledge laid out in the standards. This finding lends some support to the common criticism that standardized testing does not test complex thinking and narrows the standards by testing a small part of the content

domain. Armed with the results of this alignment research, these states could accurately address these issues in their assessment design. This study also demonstrated that each of the four criteria measured different aspects of alignment.

Another result of this study was a better understanding of the participant role. This study used six participants (Webb recommends using at least three participants). Webb (1999) noted that the participants could have benefited from more training at the beginning of the process. Some participants wanted to code near matches instead of exact matches and this confused the analysis. The participants needed more guidance about making distinctions relative to the depth-of-knowledge criteria and more explicit guidance about how to match an item to more than one standard based on the central content of an item. Webb (1999) also found that it could be helpful to put the standards in context so the participants know each state' s purpose for the standards and how they were created. During the review process the participants focused purely on the objective-item match and did not have an opportunity to critique the quality of each component and Webb (1999) found that the participants were frustrated by this constraint. While it is important to stay focused on the task at hand, it could be helpful to gather this feedback throughout the process as a means to inform future standard or assessment development work.

Webb (1999) concluded that tradeoffs between these four alignment variables are realistic but it is important to look at broader approaches to assessment to understand how other aspects of education (e.g. those discussed in the general Webb methodology but not specifically studied in his alignment process) complement the process. Unfortunately, these aspects are harder to measure and have not been included

in a formal alignment study. One limitation of Webb' smethodology was that the range of knowledge criterion did not look at the breadth of the measured objective in terms of how many different ideas are combined under one objective. If an objective were very broadly stated it was still considered assessed if it had an item matched to it, regardless of what else within that objective was not assessed. With objectives that combine many different ideas with possibly different cognitive expectations, it was easier to satisfy the range-of-knowledge criterion but this may result in a lower depth-of-knowledge result as the complexity of the objective might have increased. The interplay between the alignment variables illustrates the benefit of using the alignment results to inform the development of both standards and assessments. Furthermore, the knowledge of these alignment criteria is being used to guide item development to ensure items meet a cognitive requirement and address a range of objectives within each standard. Another limitation with the Webb methodology was that it did not capture the fact that assessments may purposefully contain items to measure standards from more than one grade. This misalignment by design should be carefully detailed in the alignment process.

In looking at the alignment study process, Webb (1999) developed a number of recommendations. If the goal were to analyze standards from more than one state, Webb recommended starting with the most detailed state standards. It would also be helpful to repeat the alignment study over time to capture the changing content of the assessment and how this may or may not impact the alignment results. Additionally, Webb has recently noted (Webb, Herman, & Webb, 2006) that averaging participants' ratings across standards and objectives might mask the different views of what the item

is truly measuring and inflate the degree of alignment across the four dimensions. Recent studies (Herman, Webb, & Zuniga, 2005; Webb, Herman, & Webb, 2006) have examined setting a minimum participant agreement requirement at the standard and/or objective level as to what the item is measuring, but this analysis is still ongoing. The Webb alignment dimensions have also recently been applied to the issue of vertical scaling. Wise and Alt (2005) discussed the possible steps to vertically align content standards and then apply the Webb dimensions to examine how the standards address the skills across the grade levels. Wise, Zhang, Winter, Taylor, & Becker (2005) provide further in depth guidance about how the vertical alignment analysis could work in terms of types of judges, types of ratings, and how the ratings could be analyzed and reports. This is an interesting extension of the alignment discussion but is still in its early stages.

Overall, the Webb model is comprehensive and provides a point of reference for the next two models reviewed. The strength of this model is its comprehensive analysis of the objective level detail, its view of alignment through four different dimensions, and the clear guidelines for what serves as acceptable levels of alignment. Another positive aspect of Webb' s work is its recognition of a broader set of issues (e.g., articulation across grades, fairness, and pedagogical implications), even though measurement of these issues is not yet fully developed. Sample reports for the Webb methodology can be found in the Web Alignment Tool Training Manual (Webb, Alt, Ely, & Vesperman, 2005). The results of a study using the Webb approach would illustrate the relationship between what is being asked of the students, how that is being assessed, and what trade-offs are made in the process.

Achieve Methodology

The Achieve methodology produces a qualitative and quantitative alignment comparison of a state' s assessment tool to the state' s standards. Rothman, Slattery, Vranek, and Resnick (2002) laid out the components of the Achieve methodology, which is designed to judge the quality of the overall assessment, as well as the individual items that comprise the assessment. Since that time, Achieve' s protocol has been further refined. Like the Webb methodology, Achieve uses a more complex model than a traditional content validity approach. The Achieve methodology also uses content experts to rate the degree of match between the standards and the assessment based on specific criteria. Unlike the Webb methodology, however, the SMEs are hired by the Achieve organization and the alignment analysis is provided as a service to the state that hires Achieve. The six criteria in this methodology build on those outlined in the Webb methodology and include: content centrality, performance centrality, cognitive demand, challenge, balance, and range. These terms will be defined and explained in more detail below. This methodology also matches items to objectives and reports findings by standard.

Like the Webb methodology, the Achieve methodology compares individual items on an assessment to the related objectives examines the degree of content and performance match, as well as the cognitive demand of the items, as compared to that stipulated in the objectives. This methodology then goes further than just looking at individual items to also consider qualitatively how sets of items matched to a standard function as a group. The participants will write summary reports about the overall items and the balance within the assessment. While potentially more time consuming

than other approaches, these additional criteria provide a more through understanding of the degree of alignment.

The Achieve methodology is applied in two stages. Each stage is briefly summarized here and then will be described in more detail below. The first stage is an item-by-item analysis to confirm the test blueprint, determine the content and performance "centrality" of each itemcompared to the objective to which it is matched, evaluate the source of challenge, and determine the level of cognitive demand. The second stage is a holistic evaluation of a set of items matched to an overarching standard in terms of the overall level of challenge, the balance and the range. The stages and steps within each stage are detailed below.

Stage 1 - Confirmation of the test blueprint. The first stage in the Achieve method focuses on item level detail only and starts with a confirmation of the test blueprint. Items are compared to the objectives, defined at the most detailed level of outcome to ensure that every item is matched to at least one objective. A match between the test blueprint and the item requires only that the item address the *same* content; the level of cognitive demand of the associated objective is not considered. Items that are mapped inappropriately are re-assigned to a more closely related objective, while items that do not match a standard or objective are eliminated from further analysis. Where a state lacks a test blueprint or the blueprint does not allow for fruitful application of the protocol, Achieve constructs a blueprint. In these instances, Achieve provides a brief rationale and communicates the findings to the state. Achieve scrutinizes the test blueprint because of its importance in developing score reports. This level of analysis is missing in the Webb approach.

Each item can have a primary and a secondary match to the objectives. The primary match is used in judging content and performance centrality, source of challenge, and level of cognitive demand (described below). The secondary match is taken into account in evaluating level of challenge, balance, and range. The use of a secondary match is similar to the Webb method where items could be mapped to more than one objective, but this model is more explicit about the degree of match and how it can be used in the alignment process. After the test blueprint has been confirmed, the participants delve deeper into the actual content of the item and how it specifically relates to the identified objective.

Content centrality. To judge content centrality, SMEs rate each item based on the degree of content match between the item and the objective it is measuring. The rating system uses a four-point scale where a "2" is a clearly consistent content match; "1A" is a match where the degree of alignment is unclear (generally because the standard is too broad to conclude that the item is clearly consistent with the objective); "1B" is a somewhat consistent match in that the item assesses only part of a compound objective; and "0" signifies an inconsistent match. This rating dimension addresses a limitation of the Webb (1999) study where a broadly stated objective may be considered adequately measured even if the item only addressed a part of the standard.

Performance centrality. In considering performance centrality, the Achieve protocol focuses on the quality of the match between the performance called for in the item and the performance described by the objective the item is intended to measure. This is similar to Webb' s(1997) method, but in the Webb approach the cognitive level of the objectives is coded in the beginning and the performance rating is made

simultaneously with the content rating. The Webb method might be more efficient, but the Achieve method allows the participants to focus on each aspect of the process in isolation. The performance centrality rating process calls participants' attention to the verbs in the objectives as compared to what the items actually demands of the student. The same 2, 1A, 1B, 0 scoring system is used for this dimension.

Source of challenge. Source of challenge is measured to ensure that items are fairly constructed and not designed to trick students. The items are reviewed to ensure they are not technically flawed (from a content perspective and by reviewing results from item analyses). For example, mathematical items are reviewed to ensure the reading level is appropriate for the grade level of the assessment and unnecessary reading is not required, while reading items are examined to ensure they measure comprehension and not prior knowledge. Reading passages are reviewed to ensure that the vocabulary, sentence structure, literary techniques, plot line, and organizational structure are all appropriate based on the grade level of the assessment. Writing prompts are similarly reviewed for accessibility, appropriate vocabulary, clarity of purpose and audience, and inclusion of basic criteria by which the sample will be scored. Each assessment item is scored as 1 for an appropriate source of challenge and 0 for an inappropriate source of challenge. If the item received a 0 for content and performance centrality then it would receive a 0 for source of challenge, as it is not a good measure of that standard. Webb recently included Source of Challenge as one of his alignment dimensions, although it is captured only through participant comments (Webb, Alt, Ely, & Vesperman, 2005).

Level of cognitive demand. A refinement in the Achieve methodology found it necessary to more precisely track the level of cognitive demand required by items to better inform SMEs evaluation of level of challenge (Slattery, 2006). Level of cognitive demand is concerned with the type and level of thinking required by students to respond to an item. The level of demand can stem from the nature of the concept assessed (some concepts are more readily understood than others) or from the kind of thinking required to arrive at a response (an item may demand routine or concrete thinking as opposed to complex reasoning or abstract thinking.) SMEs formally rate each item on a scale ranging from Level 1 (recall or basic comprehension) to Level 4 (extended analysis, typically over an extended period of time). Level 4 items are not usually found on large-scale, on-demand tests. The next stage in the application of the Achieve protocol shifts from a focus on individual items aligned to objectives to sets of related items aligned to a larger standard.

Stage 2 - Level of challenge. Level of challenge is a global judgment (not item specific) that qualitatively captures whether the collection of items mapped to a given overarching standard appropriately challenges students in a given grade level. Ideally, items within each standard should range from simple to more complex. SMEs provide a brief written evaluation of the level of challenge for each set of items tied to a specific standard, describing how the "overall demand" compares to thatexpressed in the standard, basing their judgment, in part, on the level of cognitive demand scores previously assigned to individual items in the set. SMEs look to see if a set of items are skewed toward one level of demand, if they are focused only on the more demanding or least demanding objectives within a standard and, where there are compound objectives,

if the items are skewed toward the most or least demanding part of the overall standard. The next step of the Achieve methodology examines the balance and range of sets of items relative to the expectations expressed in the standards.

Balance. Balance, like level of challenge, is a holistic evaluation. It looks at a set of items mapped to a given standard to determine how closely the set of items measures the breadth and depth of the content and performances expressed in the related standard. The relative importance the test items give to content and skills should be proportionately similar to what is stated in the standards. The SMEs comment on objectives within a standard that are over or under- assessed, redundant items, and how the overall set of items measures content they think is important for that level. The analysis allows the experts to focus on how they view the balance of the assessment as compared to the standards (Rothman, 2003). Again, this is captured qualitatively and builds on the expert knowledge of the SMEs, which is similar to Webb' s(1997) balance criterion, although that measure is quantitative. Webb' s balance calculation only determines if the objectives are equally represented, but that might not be meaningful if one area of the standards should be emphasized more through the assessment (Rothman, 2003). The quantitative measure facilitates comparison across states or districts, while the qualitative measure provides information more informative to the standards and/or assessment revision process.

Range. The range criterion also considers a set of items matched to a standard, but it measures the standard coverage. Range is a quantitative measure of the proportion of the objectives within a standard that are measured by at least one item. Ranges between 0.50 and 0.66 are acceptable and above 0.67 is considered good

coverage. This is similar to Webb' s(1997) range calculation although his methodology uses 50% coverage criterion. It is possible for a test to be well balanced, but have low coverage (and vice versa) and so it is important to consider both of these criteria.

At the close of the alignment review, SMEs look across all of the over-arching standards (i.e., at the assessment as a whole) to determine the overall rigor of the assessment and how closely it succeeds in measuring the content and performance described by the standards. Achieve then produces a comprehensive, technical report to the state that is kept secure because it contains detailed commentary on actual test items, and a policy level report meant for the state to release publicly. Sample policy alignment reports can be found at www.Achieve.org.

Methodology application. Rothman et al. (2002) applied the Achieve methodology to the evaluation of alignment between assessment and content standards in five states. The process began with a training of expert participants. The participants represented a diversity of viewpoints and included classroom teachers, curriculum specialists, and content experts. They were trained through the use of anchor items to illustrate each of the rating criteria. The state standards were a crucial starting point of the alignment study and therefore the assessments were only as good as the standards to which they were mapped.

Rothman et al. (2002) found that states with standards written in global terms received low ratings as it was more difficult to determine accurate item-standard matches. Overall this study did find that the assessment items were well matched to content and performance standards. Most states also fared well with respect to the source of challenge criterion. However, this study found that the states were not doing

a sufficient job of assessing the full range of standards and objectives, and the most challenging standards and objectives were under sampled or omitted (similar to Webb, 1999). With respect to balance, they found that the sets of items were too focused on the less important standards, a finding that was also supported by the level of challenge results.

Rothman et al. (2002) emphasized the need to focus on the issues of balance and challenge in the design and selection of state assessments. Their study illustrated both the drawbacks and strengths of the Achieve alignment method—the process can be very time consuming and expensive to undertake, but it can result in a thorough understanding of the strengths and weaknesses of a state' s assessment system.

Surveys of Enacted Curriculum (SEC) Methodology

While many teachers may think they are assessing what is taught and vice versa, assessments present different stimulus conditions than that used in the classroom and teaching and assessing are often "institutionally dichotomized" (Cohen, 1987). Porter and Smithson (2001) developed the SEC alignment methodology to help people involved in the education process see the connection between what is taught in the classroom and what is assessed, and they applied it in 11 states and four urban districts[4]. This methodology was developed to quantitatively compare degrees of alignment for standards, assessments, and instruction across schools and states. The SEC methodology builds on a content validity approach, but also measures the instructional content purportedly taught and captures this information at both a detailed and more general level of analysis.

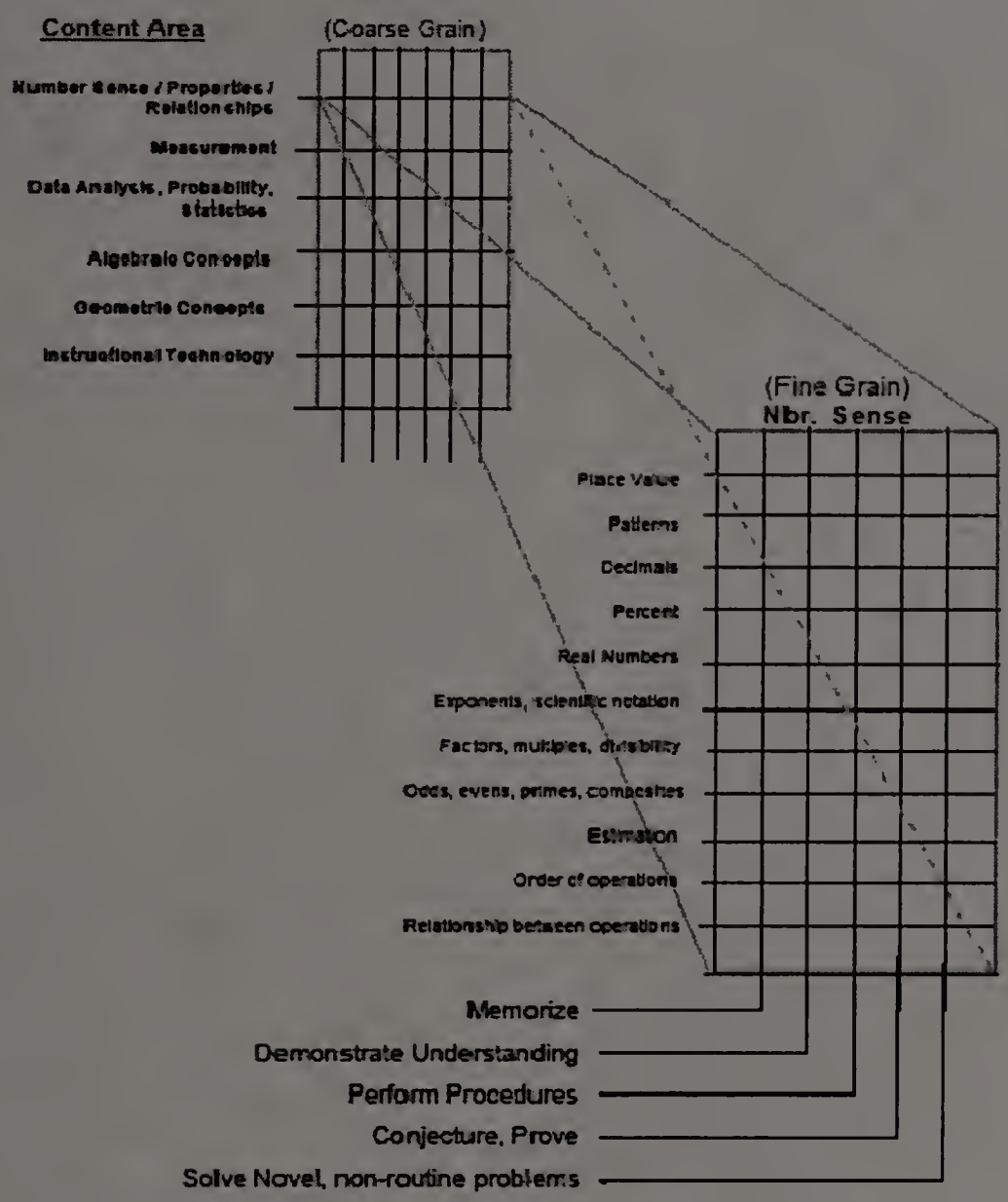
---

[4] Development and application of this model were supported by the Council of Chief State School Officers (CCSSO) through grants from the National Science Foundation and a state collaborative project.

The SEC alignment methodology comprises alignment analyses of standards, assessments, and instruction by use of a common content matrix or template that allows comparison across schools, districts or states. The SEC methodology begins with a coding process where the content and cognitive levels are determined for the standards, the assessment items, and the instructional focus. The frameworks are coded at the smallest unit possible. Coding at the objective level is similar to the Webb and Achieve methods as the results can then be summarized and reported at the strand level (e.g. Number Sense and Geometry, sometimes called standards or content areas). The assessments are coded at the individual item level. Content experts, teachers, and people familiar with the frameworks code both the standards and the assessments. Instruction content is coded at the classroom focus level and this will be discussed in more detail later in the chapter. There are three main alignment dimensions in the SEC methodology: content match, expectations for student performance, and instructional content. These dimensions are discussed below and then an application of the SEC methodology is reviewed.

Content match. The SEC method employs a content matrix of two dimensions: content topic and cognitive complexity/demand (CCSSO, 2002). The task for SMEs is to review items and match them to the topic and complexity cells in the matrix. An example of a content matrix is presented in Figure 1.

Figure 1 - Example of SEC Content Matrices
(Porter & Smithson, 2002)



The SEC methodology has developed specific topic descriptors for elementary, middle, and high school. One area of criticism of this methodology is that the number of content areas can be difficult to manage. However, the benefit is an exhaustive common view of all the content in each area of the educational process. The topics can also be reported at a fine or coarse grain level as shown in Figure 1. The fine grain level displays all of the topics by cognitive area and the coarse grain level rolls up the results to the six broad topic areas, which are similar to strands of content (e.g. Number Sense and Patterns). Thus, the method provides information similar to that gained from

traditional content validity studies, but also information at a more micro level that is more likely to better inform instructional and curricular changes (Porter & Smithson, 2002).

Expectations for student performance. The items, standards, and instruction are also coded based on expectations for student performance. This measure is similar to the depth criterion in the Webb approach and performance centrality measure used in the Achieve model. The SEC method utilizes six levels of cognitive demand or expectations for student performance. These areas are: memorize facts, perform procedures, demonstrate understanding, conjecture generalize prove, solve nonroutine problems, and make connections. These terms were chosen to be more behaviorally oriented and indicate knowledge and skills required of students as a way to help teachers to describe the cognitive expectations they hold for students (Porter & Smithson, 2001). Porter and Smithson recommend using the same cognitive levels for each area of analysis as a means to accurately make comparisons across the instructional content, standards, and assessments.

While the terms and their definitions differ across Webb, Achieve, and SEC methods, all three approaches highlight the difficulty in training the expert participants to understand the distinctions between the cognitive levels. The cognitive areas, however, are an important part of the alignment process to address the criticism that standardized tests are "dumbing down" thestandards. Through an examination of the match between the cognitive demand in each of the educational components, assessment items, standards, and instruction, the alignment measure can accurately reflect where differences appear as a means to address the issue of less challenging

curricula. The common mapping language allows the alignment results to illustrate comparisons of classroom practice to standards and assessments as well as comparisons among states, districts, and individual teachers.

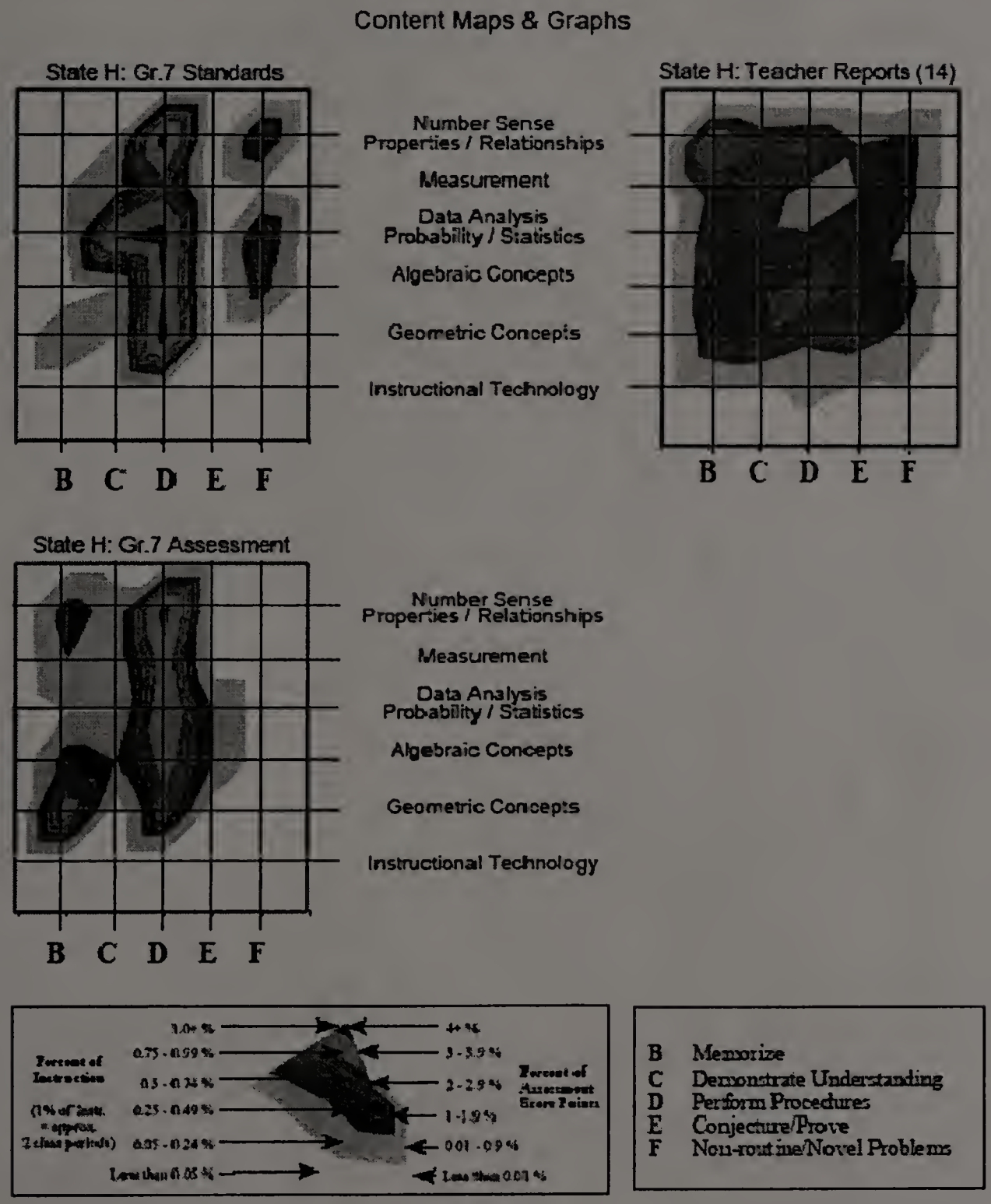
Instructional content. Unlike the other two alignment methodologies, the SEC includes a measure of the instructional content. Porter and Smithson (2002) emphasized the importance of including an instructional content component because it serves as an intervening variable when looking at student achievement gains due to standards-based reform. Through surveys, teachers code the instructional content as they think about a pre-selected target class over a specified period of time. Then, the teachers estimate the emphasis allotted to that topic for each of the cognitive areas. This is then summed to determine each topics proportion of total instructional time (Porter, 2002).

The SEC methodology provides a snapshot of practice over a period of time, which is useful in answering the question to what extent is teaching reflective of standards and assessments (Blank et al., 2001). This is a critical question that is not directly addressed by the two other alignment approaches. The benefit of the survey approach is that it allows data collection from a large number of respondents and is relatively inexpensive. Other data collection approaches such as daily logs or classroom observations will be more expensive, time consuming, and intrusive on the classroom. Porter (2002) acknowledged the weaknesses of the SEC approach in that findings are limited to what is asked, it can be subject to self-report bias, and it may be difficult to capture the complexity of instructional practice. Nevertheless, the survey tool has been piloted in multiple settings (Blank et al., 2001) and has proven useful to

address the many questions educators and policymakers have about patterns and differences in standards and instructional practices across classrooms, schools, districts, and states.

The result of the SEC coding across standards, assessments, and instructional content is that each cell in the two dimensional matrix (content by performance expectations) represents the proportion of content, assessment, or standards in that cell and these three pieces can then be compared to determine the degree of alignment. Each area matrix is compared to another to determine the degree of alignment. This resulting alignment index is: $1-\left[\left(\sum|X-Y|\right)/2\right]$ where X is the cell proportions in one matrix and Y is the cell proportions in the other (Porter, 2002). The values range from 0.0 to 1.0 and the index is the sum of cell-by-cell intersects. The results are presented on topographical map layouts to show the relative areas of concentration and facilitate easier comparisons. An example of a topographical map is presented in Figure 2. The results of an SEC alignment analysis illustrate gaps in the assessment, the curriculum, or the instruction, which can then be used to guide additional discussions about what, if any, steps need to be taken to address these gaps.

Figure 2 - Example of an SEC "Topographical" (Content) Map (Porter & Smithson, 2002)



Methodology application. Blank, Porter, and Smithson (2001) studied the degree of alignment between instruction and assessments across six states using the SEC approach. As with other alignment approaches, the participant role was crucial to this process. Specialists were brought together for a two-day workshop to code the

assessment items and standards. At least four raters independently coded each test analyzed. Because one assessment item could potentially assess different areas of content, this procedure limited raters to matching each item to up to three topic areas by student expectation combinations. To capture the instructional content piece, 600 teachers from 200 schools across six states completed the surveys in eighth grade mathematics.

The results indicated that the alignment of assessment and instruction within a state was similar to the alignment of assessments across states. That is, the alignment indices derived from cross-state comparisons of tests and standards were similar to those indices derived for comparisons of tests and standards within a state. Alignment of the state assessments to National Assessment of Educational Progress (NAEP) Grade 8 math and reading assessments were also conducted, and they found there was slightly higher alignment between state assessments and instruction within the state than there was between instruction within the state and NAEP. On the zero to one alignment index scale, across the six states the average alignment among state instruction and state assessment ranged from .23 (grade 8 science) to .42 (grade 4 math), and the average alignment between state instruction and the NAEP assessment ranged from .14 (grade 8 science) to .41 (grade 4 math). However, it should be noted that this study was conducted pre-NCLB and none of the states studied had high-stakes attached to the assessments (which would probably influence the degree to which the assessments influence classroom instruction).

The involvement of teachers in the data collection process for the SEC methodology means the alignment process itself as well as the accompanying results

can directly impact the teachers. The SEC methodology is one way to get inside the "black box" of classroom instruction and examine these practices in the context of a large-scale study, which is necessary to evaluate the effectiveness of any reform initiative (Blank et al., 2001). To gain teachers' participation in SEC studies it is imperative that it be voluntary and the results not be tied to any accountability measures. Additionally, teachers should be given individualized results and provided with training about how to use the results (Blank et al., 2001). Results of SEC studies have been used as the basis for professional development opportunities using the in-depth curriculum data for improving instruction in math and science (Blank, 2004).

An SEC methodology based approach to professional development draws on what effective professional development should be: linked to content standards and emphasizing active learning, focused on continued improvement using data and formative evaluation, building school-based collaboration and networking to share teaching ideas and strategies for improvement (Blank, 2004). The results from an SEC study can help teachers to visually see the areas of the standards that are not being taught or taught with only limited time, require a higher level approach, and where there are gaps. The data can be presented as specific instructional topics and this level of analysis is important for teachers to then address the results in their classrooms (Blank, 2004).

A student survey would be an interesting addition to the SEC methodology as a way to understand their view of the curriculum they see in the classroom. While this has been noted as an option (Porter & Smithson, 2001), the concern is whether students will be able to accurately represent what they have learned – if students are confused,

they will not be able to express what was taught. Even so, at some point it could be interesting to gather this information and combine it with the results to see if it contributes additional information when looking across the class and comparing it to the teacher survey results. A positive outcome of student involvement in alignment research would be students' increased ability to selfassess and gain ownership of their learning, which has been noted as an important practice for effective learning (Leahy, Lyon, Thompson, & Wiliam, 2005; McTighe & O'Connor, 2005). Blank et al. (2001) did collect student data from 123 classrooms. Correlations were computed between student and teacher responses in order to determine degree of consistency between student and teacher reports. Student data were aggregated by class so that comparisons could be made between the teacher reports and the class average from student reports. Within the math area, student and teacher reports correlated well. Of the 49 survey items all but three had significant and positive correlations.

Porter (2002) summarized the multiple benefits of implementing an SEC approach to alignment. It is an efficient process, once both the coders of the assessment and standards and the teachers being surveyed have been trained. The process allows for an objective evaluation of the alignment goals. The result of this study is a quantitative measure of alignment that can be used to examine the effect of reform policies over time. Since this approach maps the education pieces to a common language and then compares the results, the process can be used to compare findings across schools, districts, and states. Additionally, teachers are an integral component of this methodology and they see the results of their involvement in a way that can

meaningfully impact their approach to instruction. It helps teachers to understand how what happens in the classroom relates to the bigger picture.

Porter and Smithson (2001) also discussed the potential benefit of developing electronic instrumentation to facilitate the data collection and analysis process. Additionally, teachers could receive immediate feedback, a profile of their own practice, summary of results of the other teachers in their district, state, or nationally, and content maps for various assessment instruments.

The SEC approach has similar limitations to Webb and Achieve. The process begins with the state standards and is only as good as what they are working from. Additionally, if the standards are not specific enough it will not be possible to tightly align the assessments (Porter, 2002). This methodology does not have the more detailed criteria beyond content and depth match, which are found in the Webb and Achieve models, and so the methodology is unable to quantify the detailed reasons behind limited alignment. Also, research is needed to understand the degree to which teachers and policy makers understand the concept maps that characterize instructional coverage. In the SEC method there may also be concern about the reliability of the content and instructional analyses given that it is survey data based on recollections over a period of time (Porter, 2002).

The survey process can also be somewhat complex for teachers given the multiple ways they are coding their instruction (Anderson, 2002). Although the two studies applying this approach had a 75% response rate (Porter, 2002), the survey response rates can be dependent on how the survey is administered. Blank et al. (2001) found that the worst response rates were seen in those schools where teachers were

given the surveys to complete on their own at their convenience and the best response rates came from those schools where the teachers gathered as a group to complete the surveys. Response rates were also higher where teachers were compensated or given professional development credit for the time it took to complete the survey. Blank et al. (2001) concluded that teachers must perceive some personal value to the information they provide. It was important that the information was confidential and that teachers were provided with individual reports if requested, while ensuring the results would not be used for teacher accountability.

Summary of Alignment Methodologies

Bhola et al. (2003) did an interesting overview of different alignment approaches and classified each according to the degree of complexity entailed in the model. Low complexity models defined alignment as the extent to which the items in a test match relevant content standards (or test specifications) as judged by content experts rating the degree of match with Likert scale ratings. This is the approach taken in more traditional content validity-type studies (Buckendahl, Plake, Impara, & Irwin, 2000; Sireci, 1998b). In moderate complexity models, content experts decide matches both from content and cognitive perspective and the result may be a reduction in the number of matches because of this additional constraint. This is the approach used in SEC where the standards, assessments, and instruction are aligned. High complexity models tie in additional criteria to give a broader view of the alignment. Webb' s(1999) approach and the Achieve approach (Rothman et al., 2002) are examples of this level of detail.

Similarities and differences across methods. The Webb, Achieve, and SEC alignment methods have not yet all been applied in a single study and so the differential utility of the results they provide cannot be accurately described. However, Appendix A provides a description of the major aspects of each method, organized by five dimensions: content, cognitive, breadth, distribution, and item quality. Based on the five criteria applied, one can see what aspects of each alignment methodology are strong or weak. The Webb approach provides the most detailed quantified results. The Achieve methodology builds on the Webb methodology with the addition of the source and level of challenge dimensions. These dimensions are a means to capture item and standard quality, which was a limitation in Webb' s method However, the most recent applications of Webb' s methodology now include a Source of Challenge criterion (Webb, Alt, Ely, Cormier, & Vesperman, 2005). The Achieve methodology also provides more qualitative information about the alignment overall and the quality of the matches. This latter point is missing in the Webb approach where an item-objective match does not convey if the objective is only partially assessed or too vague to be assessed. In this way the specific coding in the Achieve methodology provides a bit more helpful information in terms of possible changes a state might undertake. The broader qualitative results from the Achieve method are helpful for a specific state application but might become cumbersome if used for comparison purposes among states. The SEC methodology is the only method that considers the instructional piece of the educational process and allows for easy comparison of assessments, standards, and instruction across states, districts, and schools. It may also be particularly useful for studying the consequences of a testing program if comparisons are conducted and

compared over time. This approach, however, does not probe as deeply as the other two methodologies do into the quality of the alignment. Thus, these alignment methods each have a different focus and each has strengths and limitations in specific situations.

Importance of subject matter experts. All of the alignment methods depend on expert participants to rate the different components of an alignment study. In selecting participants, all approaches emphasize the importance of knowledgeable experts who are familiar with the standards, assessments, and instructional components. It is also critical that the participants are familiar with the knowledge and skill levels of the tested population (Sireci, 1998b).

Using expert participants is an important part of the process as studies have shown test publisher ratings may differ significantly from expert participants (Buckendahl et al., 2000). Additionally, the participants may be influenced by the fact that they are told the categories that the items, standards, or instructional content must fit into and are constrained by these definitions (Sireci, 1998b). Furthermore, the participants can be influenced by social desirability of what they think is expected, leniency to find a match, and guessing (Sireci, 1998b).

Regardless of the alignment method employed, it is important that the level of SME agreement is reported. Rothman (2003) discusses the varying levels of participant agreement among the different types of studies. While Achieve uses SMEs that are highly trained in the Achieve methodology, the Webb and SEC methods appear to have more limited training. However, the Webb and SEC alignment results quantify the levels of participant agreement. The Webb methodology provides explicit details about the calculations used to capture the reliability of both the cognitive level coding and the

item-objective matches (Webb, Alt, Ely, & Vesperman, 2005). The SEC method also computes inter-rater agreement levels. Webb et al. (Webb, Alt, Ely, Cormier, & Vesperman, 2005) noted the importance of having an adequate number of participants to ensure the reliability of the coding. Earlier iterations of the Webb methodology recommended three to eight participants but Webb now finds that ideally more than six, but anywhere from five to 12 participants, is better to ensure a greater degree of reliability in the coding. While some guidance is provided as to acceptable levels of agreement, this calculation in general serves as a check as to the reliability of the expert judgments.

Challenges in evaluating alignment. Alignment research can be difficult to conduct for six main reasons. First, not everything that is in the standards can be assessed through large-scale standardized assessments. Webb supported broadly defined assessments to include classroom, district, and statewide assessments so as to capture a broader view (Webb, 1997). However, in the alignment studies reviewed this comprehensiveness does not seem practical. All of the alignment studies used statewide, standardized assessments as their comparison and this seems most in line with the expectations laid out in NCLB. Second, standards may be written at multiple levels and tests may be written to align with standards at the highest level, but the alignment study may use a more detailed level for the standard comparison (Ananda, 2003a). Third, standards may be written to different levels of specificity and may be written so generally that many different types of content are incorporated so determining a match is difficult (Rothman et al., 2002). Fourth, the terms within the standards may have multiple meanings to different people. Webb (1997) provided an

example with the phrase "demonstrate a range of strategies" and discussed how this was difficult to interpret and therefore assess. Fifth, items may measure multiple content standards, which can result in error among expert judgments (Le Marca et al., 2000). Sixth, some standards may not be easily assessed, may be redundant within a level, or tests may be designed to assess multiple grade levels. For these reasons perfect alignment will not be expected (Ananda, 2003a) but most alignment methodologies discuss levels of acceptable alignment for the different dimensions.

Given the range of criteria used in an alignment study, states need to be clear about their alignment goals. For example, some states might not value the goal of the assessments having a balanced distribution of items across objectives within a standard and may want greater emphasis within specific areas (Ananda, 2003b). Most states will want to ensure their tests adequately measure the intended strands or objectives, and so a traditional content validity study that focuses on this congruence, or the dimensions of alignment models that look at this congruence, may suffice. These three methods of alignment offer a range of approaches but each method will result in a deeper understanding of how well a consistent message is sent to students about what is valued.

Alignment as a Form of Professional Development

Even if the tests are aligned to a solid curriculum framework, the teachers are still the gate keepers through which the students receive the content. Crocker (2003) noted this fact and used this to support her view of increased teacher involvement in the testing and alignment process. Alignment research can demonstrate how the tests support what is supposed to or is being taught in the classroom. However, teachers' involvement in the alignment process can be a powerful means to help the teachers

understand how the alignment of assessments and standards can support what happens in the classroom. The value of applying an alignment methodology is both the end result and the process that involves teachers moving from a focus on the textbook to a dynamic focus on student learning with an understanding of where students are headed and how the teacher can help to get them there (McGehee & Griffith, 2001). This next section details the use of alignment results as a form of data-based decision making and the use of alignment studies as effective forms of professional development.

Data-Based Decision Making

Results of large-scale assessments provide individual student data, which allows for instructional decisions to be data-driven. This has been applied rigorously in the K-12 arena. Through the use of released items and alignment processes such as those discussed earlier, the assessments themselves may become agents of educational change. McGehee and Griffith (2001) studied whether professional development projects focused on the alignment of written and taught curriculum with criterion tests could work to improve student achievement. The authors found teachers' perceptions of the test to be critical for how the test will be used. If the test was perceived as good and innovative teachers will examine their own practice in light of what the test asks. To positively influence classroom practice it is critical that information is released about the test item content and how the test was constructed, that the test aligns with the standards, and that teachers are involved in the alignment process (McGehee & Griffith, 2001). Satisfying these requirements helps the teachers to gain confidence in the test and how it relates to what needs to be done in their classrooms.

Porter and his colleagues are currently involved in using the results of an alignment study as a key component of professional development focused on data-based decision making. Teaching practices can be improved based on discussion and analysis of data (Love, 2000), and teachers can work together to improve consistency of content (Blank et al., 2001). The value of discussions focused on assessment results, how the results relate to the standards, and what this relationship means for teaching practice can be a powerful means to unite a school and initiate school reform (Martone, 2005). When discussion is structured more systematically around the alignment process one can imagine it would be even more effective. The results of an SEC alignment can illustrate differences in practice among teachers, determine whether those differences need to be addressed, and start the conversations about what additional supports the teachers need (Porter, 2002). The teachers can learn how to read and analyze the data and then progress to discussions about how to use the results to inform their practice (Blank, 2004).

Alignment Research as an Approach to Professional Development

Research on effective forms of professional development illustrates insights into what additional characteristics could be helpful in an alignment study. It is critical that professional development focuses on what works, is curriculum relevant, and is results oriented (Cizek, 2001). Alignment research addresses the last two points and can "hook" teachers to more deeply discuss their practice and focus on what works. Involving teachers in alignment research helps them to learn through application about assessments and how it relates to instruction and standards and this link is a critical link for educational improvement (Guskey, 2005). Teachers' involvement in he training

and subsequent scoring of tasks makes them more reflective, deliberate, and critical of their own classroom instruction and assessment (Cizek, 2001). This would translate to the tasks required in an alignment study as teachers would need to think deeply about the content of the assessment compared to the standards and their own instruction. Furthermore, teachers' involvement in the assessment process, through creation, review, scoring, or aligning, can help to change teachers' attitudes and approaches to standardized testing as they begin to teach "for a test" or with an understanding of a test as opposed to teaching "to a test" which has a more automated connotation(Crocker, 2003).

Beyond just participating in an alignment study, the teachers will need other supplemental activities to help them see the alignment process as useful in their classroom. To enhance the impact of the alignment research, it would be helpful to discuss with the teachers how the alignment of the assessments, standards, and their instruction could be different from the current practice and how it can directly benefit their students (Sparks, 1988). It would also be helpful to allow for small groups to talk about the positive and negative aspects of the alignment process, hear testimonials from people who have been through this process, and learn about some of the theories and research underlying the process (Sparks, 1988). While these activities would be secondary to the alignment process, they are important components if the activity is to be used as a true form of professional development. Because the alignment process will be a hands-on activity, directly related to the content areas of teaching and what the teachers do on a daily basis, knowledge gained from this practice can be integrated into

the teachers' practices and resultin the teachers having increased knowledge and skills (Garet, Porter, Desimone, Birman, & Yoon, 2001).

A number of studies examined the effectiveness of teachers' involvement in scoring assessments as a form of professional development. In Falk and Ort' s(1997) study of teachers' involvementin the scoring of performance assessments, they found that through conversations with other teachers about the evaluation process the teachers better understood what knowledge was assessed and how it related to the standards. The teachers gained insights into children' s thinking and benefited from the collaboration with other teachers to see what is possible and what changes could be made. Collaboration is a key component of successful professional development experiences (Borko, Davinroy, Bliem, & Cumbo, 2000; Borko, Mayfield, Marion, Flexer, & Cumbo, 1997; Franke, Carpenter, Levi, & Fennema, 2001; Garet et al., 2001; Stipek, Givvin, Salmon, & Macgyvers, 1998; Swafford, Jones, Thornton, Stump, & Miller, 1999; Wolf & White, 2000). Similar results could be possible as teachers are involved in alignment research. Beyond just the coding process and the decisions that entails, teachers can have discussions about the results and what it means for the school and the teachers (Blank, 2004).

Borko, Mayfield, Marion, Flexer, & Cumbo (1997) studied teachers' involvement in creating and scoring assessments. This process involved workshops after school for 1 1/2 - 2 hours twice a week where materials were distributed and discussed, applications of earlier assessments were brought to the group and small groups shared examples of students' work and the issues they were seeing. The study found that it was important for the staff developers working with the teachers to be able

to relate the new ideas to where the teachers' needs and interests were located(Borko et al., 1997). This would be addressed in the application of the SEC alignment approach as the results would lay out the existing practice but also illustrate what are possible changes in instructional content.

When thinking about how alignment research based professional development can occur, it is important to focus that research at the teaching site and provide sufficient time for the teachers to apply what they are learning and then follow up in the group setting with questions (Borko et al., 2000). It is also important to provide the teachers with materials and resources to try different approaches (Borko, 1997). In this way the support provides the teachers with options as opposed to a mandated solution. This would be helpful in an alignment process where there is not one correct answer and different variables need to be examined.

Through No Child Left Behind, student assessments have become a dominant feature of the educational process. An important component of the effectiveness of No Child Left Behind legislation is to help teachers to use these assessments to guide their instruction to ensure student achievement gains. Teachers need to understand the value of the assessments, how the assessments relate to what they should be teaching, and how to make changes in their approach based on the results they see. Teachers' involvement in alignment research is one way to help teachers become more familiar with the assessments and the standards on which they are based. Alignment research that incorporates the findings about effective forms of professional development studies can ensure teachers apply what they are learning through the alignment process to their classroom.

Issues Specific to Adult Basic Education

The adult basic education (ABE) population has many characteristics that make it a unique population to work with concerning alignment and professional development needs. Massachusetts has invested money and resources in adult education and is considered one of the leaders in the nation for reform efforts. A report by Comings and Soricone (2005) provides a helpful case study example of the profile of the Massachusetts adult education sector. Massachusetts understood that ABE was a key component of statewide economic development through the production of a more educated workforce. As of 2002, 35% of the state' s 3.2 million workers did not speak English well, needed to obtain a GED, or needed more advanced literacy or math skills (Comings, Sum, & Uvin, 2000). From 1987 to 2002 Massachusetts increased its funding of adult basic education from $4 million to $45 million. Part of the reform efforts to improve adult basic education includes better assessment and accountability but in the short term Massachusetts had to rely on existing off-the-shelf assessments. The development of the MAPT for Math is part of Massachusetts' commitment to developing assessments that would better measure their students' growth.

The Comings and Soricone (2005) report provides a summary of the diversity of services and population that make up the adult education field in Massachusetts. Massachusetts delivers services through a variety of providers, which include: school districts, community-based organizations, institutions of higher education, and municipal agencies. Some programs offer intensive 15-20 hours per week while other programs offer 5-8 hours per week. Each program must offer at least three levels of instruction within a given instructional period. Given the variety of approaches and

learner levels, flexibility in an assessment system is important. Students come from a variety of backgrounds. As of 2002, 72% of the students were between 19 and 44, 21% were over 45, and 7% were between 16 and 18. The students are white (26%), Hispanic (23%), African American (19%), Asian American (12%), and Native American or Alaskan (1%). But for the majority of all students English is not the only language spoken in the home, although it is the most common first language. Also in 2002, 59% of the students were also employed and 25% were looking for work. This diversity within the student population illustrates the need for an assessment that spans a wide skill range and programs that offer flexibility in how services are offered.

Teachers have a wide variety of backgrounds as illustrated in the Comings and Soricone (2005) report. Almost all teachers have a 4-year college degree. As of 2002 voluntary certification as adult education teachers has been offered in Massachusetts as a way to professionalize adult education. But unfortunately teacher turnover is very high. In 2002, 57% of the teachers had been in their programs for less than 2 years and only 19% were there for more than 5 years. Part of the issue is the shortage of full time teaching positions. As of 2002, only 11% of the teaching positions were full time. Many educators work in adult education to gain experience and then move to the K-12 domain for better salaries and benefits. Recent changes in pay and support for teachers hopes to reduce this turnover. These changes include: increased pay, professionalization of adult educators through licensure programs, increased program and staff development. This study will address the issues of how teachers can become more familiar with a new assessment designed to measure the students' learning based on the statewide curriculum frameworks. In this way, the alignment process can be one

form of professional development to help teachers bring together the assessment, standards, and instructional components in a meaningful way.

Summary

Alignment is a means to understand the degree to which different components of an educational system work together to support a common goal. In this age of accountability, it is important that state organizations, districts, and schools support each other to send a consistent message to students about what is required. Alignment research is one method to demonstrate this consistency of message or to understand what changes need to be addressed to ensure every student has the opportunity to learn the content on which they are assessed, and to demonstrate his/her proficiency. Furthermore, to meet the expectations of alignment under NCLB, states will need to conduct independent analyses of the alignment between their tests and curriculum frameworks, and if any gaps are discovered, they will need to take corrective action.

All three methodologies reviewed here start with the basic evaluation of the alignment of the content and cognitive complexity of standards and assessments. The SEC methodology also includes an instructional component. On to this foundation the Webb and Achieve methodologies layer additional criteria to better understand the breadth and range of comparison between the standards and the assessments. Then the Achieve methodology also includes an overarching view of the sets of items to look at the broader quality of an assessment relative to the standards on which it is based.

When deciding between these three approached to alignment research, it is important to understand the resources available, both time and personnel, and the ultimate purpose for the research. However it is accomplished, alignment research

should be viewed as an ongoing process to continually understand how the standards, the assessment, and the instruction support each other to deliver a consistent message to students about what is expected.

Through NCLB, student assessments have become a dominant feature of the educational process. An important component of the effectiveness of NCLB is the use of assessments to improve instruction. Summative assessments, such as statewide standardized tests, are criticized for providing too little information too late in the process to help teachers implement changes in the classroom so assessments are used mainly for reporting as opposed to as opportunities for learning (Scherer, 2005). Teachers' involvement in alignment research allows the seemingly removed summative assessment to be connected more directly to the content standards and classroom instruction. Through an alignment process, teachers look closely at individual assessment items, connect these items to the standards, and think about how this translates into what they do in the classroom. Teachers can see how the assessment compares to what they are expected to teach and what they emphasize in their teaching. Teachers need to understand the value of the assessments, how the assessments relate to what they should be teaching, and how to make changes in their approach based on the results they see. Teachers' involvement in alignment research is one way to help teachers become more familiar with the assessments and the standards on which they are based. Alignment research that incorporates the findings about effective forms of professional development studies can ensure teachers apply what they are learning through the alignment process to their classroom.

Professional development activities designed to support alignment research is one means to help teachers take the knowledge they gain from an alignment study and apply it to their classroom instruction. If the goal of alignment research is to ensure that the assessments are based on the standards, teachers need to be included in the research to ensure that what is in the standards is what is taught. If the results of an alignment study are to ultimately impact student achievement, teachers are a critical conduit for what is emphasized in both the assessment and the standards and they need to be included in the assessment process. Alignment research presents an effective means to efficiently illustrate the desired connections between what is in the standards, what is assessed, and what should be taught, as well as how cognitively demanding the content needs to be.

Alignment research represents an exciting and powerful way to bring different parts of the educational system together in a systematic and efficient way. While the process may be costly, as it is dependent on expert participants and takes time, the results send a powerful message about the state of these educational components, assessments, standards, and instruction, and what might need to be addressed going forward. Chapter 3 will provide an overview of the methodology used in this specific alignment study.

CHAPTER 3

METHODOLOGY

Introduction

The purpose of this study was to explore teachers' participation in the analysis of the degree of alignment between the Mathematics Massachusetts Adult Proficiency Test (MAPT) (hereafter referred to as the MAPT for Math) and the Massachusetts Adult Basic Education Curriculum Framework for Mathematics and Numeracy (hereafter referred to as the Math ABE standards). This study explored how well the MAPT for Math was aligned to the Math ABE standards based on the criteria in the Webb alignment methodology. Additionally, this study explored how participation in an alignment study influenced the teacher participants' knowledge and attitudes about the test and the standards and how this experience influenced their approach to instruction.

This study was significant in two ways. First, I analyzed the degree of alignment between a standardized test and statewide curriculum standards for the ABE population. While alignment research on assessments and standards has occurred in the past (Blank et al., 2001; Rothman et al., 2002; Webb, 1999) it has never been systematically done in the ABE population. Furthermore, the MAPT for Math is a multi-stage computer adaptive test and applying an alignment process to this type of test has not been done before. The results of this study will specifically inform the test and standards development process for adult basic education in Massachusetts. Finally, the alignment process for this study occurred in the adult basic education community.

Alignment studies are much more common in the K-12 areas so this study will also broaden the application of alignment research.

Second, this study involved ABE teachers in the alignment process. The teachers participated in the coding of the assessment and the standards and also in a review of the alignment results. Additionally, throughout the alignment process the teachers were guided to think about how their participation in the alignment study might influence their approach to instruction. This link between assessment, standards, and instruction is critical but rarely are all three components addressed in alignment research.

I became interested in studying teachers' involvement in alignment research for several reasons. First, as a past fourth grade teacher in New York State I was faced with a high-stakes standardized assessment and statewide curriculum frameworks. Individually I worked to analyze these two components while thinking about how to incorporate what I learned into my approach to instruction. I wanted to integrate a thorough understanding of the assessment and the standards with how I structured my daily, weekly, and monthly lesson plans. My goal was for the students to see the annual standardized test as a way to show all that they had learned throughout the year rather than as something they had to fear and wonder if they were prepared for. Through this analysis process I wondered if there was a method to do the review and reflection more systematically and to collaborate with others about the process.

Through my studies and work as a doctoral student I became interested in the formal area of alignment research. Alignment research combined my interest in assessment and standards in a methodological way. I believe that including teachers in

the alignment process will increase their understanding of the assessment and the standards. Additionally, providing opportunities for the teachers to reflect and discuss how the process might influence their approach to instruction will lay the groundwork on which future professional development experiences can be built.

The purpose of this chapter is to provide information about my methodological approach: overview of the alignment components; selection of participants; methods for making contact; data gathering methods; data analysis; and the trustworthiness of the study.

Overview of the Alignment Components

This study analyzed the degree of alignment between the MAPT for Math and the Math ABE standards. An overview of these two components is important to lay the foundation for the methodology of the study. In the ABE community, students are placed at different learning levels or test levels that have grade equivalent meanings. The grade equivalencies of these levels are detailed in Table 1 below.

Table 1 – Grade Level Equivalencies for each Test Level

ABE Learning Level	Grade Level Equivalent
1	0.0 – 1.9
2	2.0 – 3.9
3	4.0 – 5.9
4	6.0 – 8.9
5	9.0 – 10.9
6	11.0 – GED

The Math ABE standards detail what skills students are expected to learn at the different learning levels of the educational process. The Math ABE standards are represented in a curriculum framework document and they are divided into four strands: Number Senses, Patterns, Functions, & Algebra, Statistics & Probability, and Geometry

& Measurement. The Math ABE standards, a 100+ page document, can be found at http://www.doe.mass.edu/acls/frameworks/mathnum.pdf. Within each strand for each learning level, objectives list the specific skills to be taught. While the content strands are consistent across the learning levels, the objectives within each standard show how the content knowledge progresses as the students move to higher educational levels.

The MAPT for Math currently assesses learning levels two through five. The assessment was not designed for level 1 students because of their lower literacy level. The MAPT for Math also does not assess the level 6 students because they take the GED as their primary assessment. The purpose of the MAPT for Math is

> to measure students' knowledge and skill in specific standards in the MA ABE Curriculum Frameworks so that their progress in meeting educational goals can be evaluated. Assuming sufficient sample sizes and test lengths, ACLS Proficiency test scores, or score gains, can be aggregated across students within adult education programs to provide a meaningful, summative measure of program effectiveness (Sireci et al., 2004, p. 2).

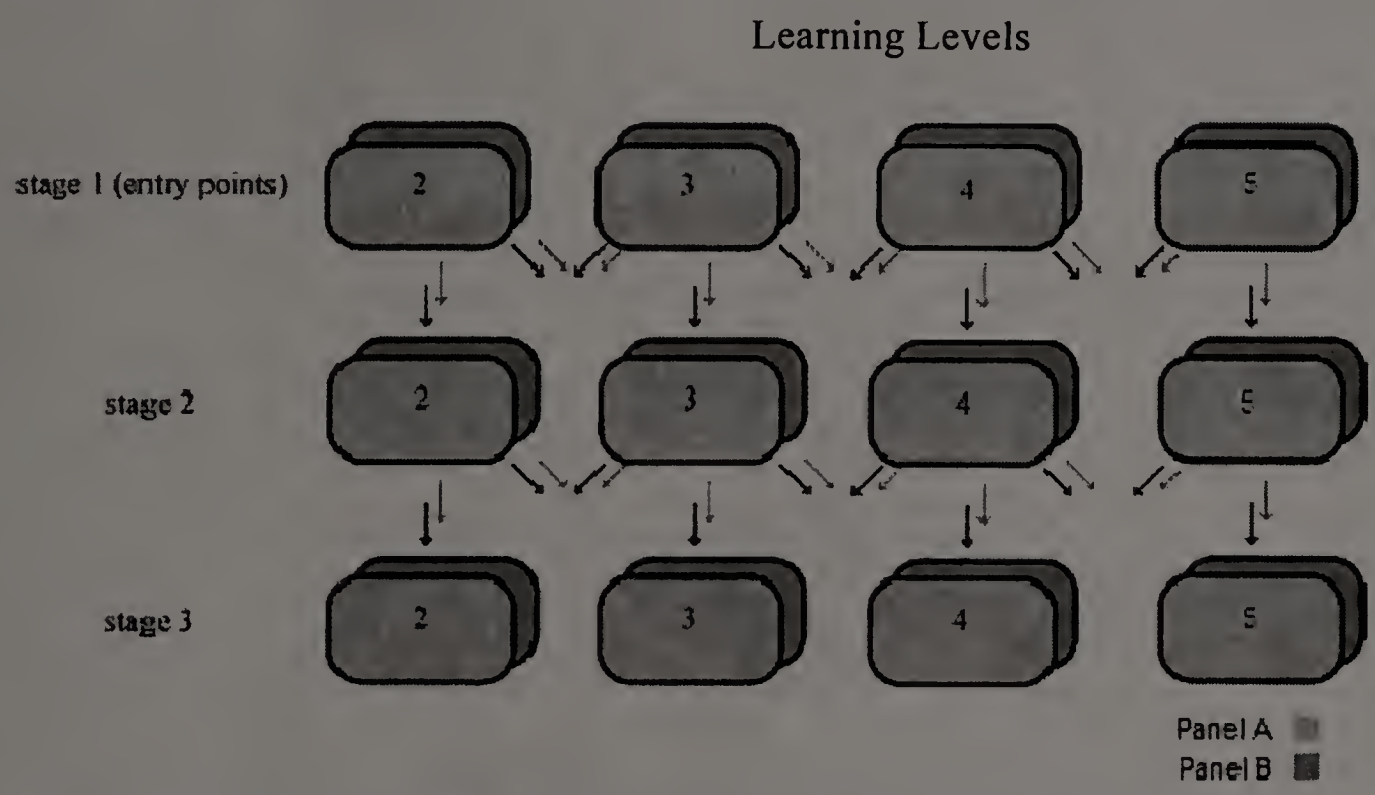
An additional purpose for the assessment is to emphasize the link between the assessment and the Math ABE standards. To accomplish these purposes the Center for Educational Assessment designed an assessment to directly link to the Math ABE standards and to assess students at levels two through five.

ABE teachers trained in item writing by the Center for Educational Assessment wrote a majority of the items for the MAPT for Math. Through this process each item was directly tied to an objective in the Math ABE standards. All items were then reviewed by a committee of ABE math experts to judge the quality of the match between the item and the associated objective on a scale of 1-6 with 6 being a perfect match. This was an initial attempt to ensure the assessment was aligned to the ABE

Math standards. The MAPT for Math only utilized items that were rated by the subject matter experts as four or above.

The operational MAPT for Math is a computer-based multistage adaptive test. The test is administered in three stages with each stage more successfully targeting the student' s trueproficiency and presenting items that will allow the most information about the student to be learned. Given the large range of proficiency in the ABE population (covering the grade level equivalency of 2.0-10.9) the design of the test is meant to avoid presenting items to students that will be too easy or too frustrating. The stages have been designed using the IRT difficulty estimates determined through the pilot testing process for each item. Students start at a level based on their instructor' s decision and complete 10 items. Based on their performance on the items in the first stage the second stage will contain items that are either easier, harder, or of equivalent difficulty (14 items). Then based on a student' s performance in the second stage the final stage will present a set of items that target the student' s trueproficiency (16 items). The overall test has 40 items. There also is a pre and post-test version, Panel A and Panel B. The illustration in Figure 3 displays the test structure design.

Figure 3 – Multi-Stage Design for the MAPT for Math (Sireci et al., 2006)



In a multi-stage adaptive test, students at the same level will not necessarily take the same test. For example, a student may start at level 3 and answer all of the items correct in the first stage. The second stage will then draw from items more representative of level 4 content. For this reason there is no one test form to use in an alignment study as there typically is with a paper and pencil test. This study simulated a straight path through the stages that a student at each level might take. This gave a representative test at each level. However, given the adaptive nature of the test, even this straight path did not ensure that only test content from the specified level appeared on that test. For example, an item might have been written to a level 3 objective, however, through the pilot testing it had a difficulty value in the range of level 4 items. This item might then appear on the reviewed level 4 test even though it is written to a level 3 content objective. In a traditional content validity study this would not be a problem but this difference in level will be highlighted through the more detailed alignment study process. This study rated the degree of alignment between a test at

each level (levels 2-5) with 40 items in each test and the Math ABE standards at that level. Tests for level 2 and 4 came from panel A and tests from levels 3 and 5 will come from panel B. This allowed a sample from each panel to be examined while still ensuring that there are alignment results for each test level.

Each overall test at each level still conformed as closely as possible to a set of underlying test specifications for that test level. The Center for Educational Assessment at the University of Massachusetts Amherst, the Massachusetts Department of Education' s Office of Adult and Community Learning Services (ACLS), and the System for Adult Basic Education Support (SABES) convened a blue ribbon committee to design the test specifications for the MAPT for Math (Sireci et al., 2004). The committee determined that the content strands from the curriculum frameworks represented the content dimension of the test. The committee also decided that the cognitive dimension of the test was best captured through a collapsed version of Bloom' s taxonomy(Bloom, 1956). The cognitive dimensions for the MAPT for Math included: Knowledge & Comprehension (KC), Application (App), and Analysis, Synthesis, & Evaluation (ASE). The committee then determined what proportion of test content needed to be represented in each area as shown in Table 2.

Table 2 – Test Specifications for the MAPT for Math

Test Level 2				
Content Strands	KC	App	ASE	Total
Number Sense	15	15	5	35%
Patterns, Functions, Algebra	6	6	3	15%
Statistics and Probability	10	10	5	25%
Geometry and Measurement	10	10	5	25%
Total	41%	41%	18%	100%
Test Level 3				
Content Strands	KC	App	ASE	Total
Number Sense	10	15	5	30%
Patterns, Functions, Algebra	10	5	5	20%
Statistics and Probability	5	15	5	25%
Geometry and Measurement	10	10	5	25%
Total	35%	45%	20%	100%
Test Level 4				
Content Strands	KC	App	ASE	Total
Number Sense	10	10	5	25%
Patterns, Functions, Algebra	5	15	5	25%
Statistics and Probability	5	10	10	25%
Geometry and Measurement	5	15	5	25%
Total	25%	50%	25%	100%
Test Level 5				
Content Strands	KC	App	ASE	Total
Number Sense	3	6	6	15%
Patterns, Functions, Algebra	5	15	10	30%
Statistics and Probability	5	10	15	30%
Geometry and Measurement	5	10	10	25%
Total	18%	41%	41%	100%

The test specifications illustrate how the proportion of content changes as the levels change. The content becomes more cognitively challenging and the proportions within the strands change as the learning levels progress. For example, students will see more Analysis, Synthesis, and Evaluation items at level 5 (41%) than at level 2 (18%). The fact that the test specifications change at each level is one of the challenges of an

adaptive test design. Now that the details of the MAPT for Math and the Math ABE standards have been discussed the next section of the methodology will review the participant selection process.

Selection of Participants

Participant selection was an important step in the research process. Based on guidance from past alignment research (Webb, 1999, 2002) six participants reviewed and rated the standards and assessments at each of the four levels of testing. Each participant rated 160 items total (40 items at 4 different levels). The six participants were all ABE Math teachers, as this is the population who will be administering the MAPT for Math and using the Math ABE standards. Within this population teachers have a range of experiences and background knowledge about the assessment and standards. There was a conscious balance between teachers who had had a high degree of involvement in the test and standard development process and math teachers who might have a more limited understanding of these components. To obtain this balance I used a purposeful participant selection strategy combined with a convenience sample (Rossman & Rallis, 2003).

Potential teachers were drawn from three sources with two teachers from each source. First, two teachers were selected from the pool of teachers who participated in the item writing and test development process. These teachers, Beth and Mary, were very familiar with how items were written to individual objectives and cognitive levels and have had exposure to the breadth of the Math ABE standards. Second, two teachers who participated in a project funded through an NSF grant were selected. This project was called Teachers Investigating Adult Numeracy (TIAN) and the goal of the project

was to help teachers and programs strengthen their capacity to provide effective mathematics instruction that is well-aligned with the state' s ABE Curriculum Framework for Mathematics and Numeracy. TIAN worked with 20 Math ABE teachers to provide in-depth analysis to two of the content strands, Number Sense and Statistics & Probability. The two teachers selected with this background experience, Judy and Len, completed in-depth work with these two strands but were less familiar with the other two strands of the curriculum frameworks. Third, two teachers were recommended by ACLS. These teachers, Melissa and Sabrina, did not directly participate in item development or the TIAN project but did have an interest in math and participated in past ACLS sponsored professional development experiences.

The six participants for this alignment study came from diverse backgrounds with a range of experiences. There were five female panelists and one male panelist. Two of the six panelists were from Western Massachusetts, Judy and Len, and the remaining four panelists were from Eastern Massachusetts. One participant, Len, had been teaching for five years and the other five participants had all been teaching for over 15 years. The participants taught students who are native English speakers and students for whom English is a second language. The students in their classes were mainly White, Hispanic, African-American, and Asian. This sample was not meant to be representative of the population.

Results from this study served as a first step in looking deeper at teachers' understanding of the connection between assessments, standards, and instruction. Future studies might be more systematic and exhaustive in the sampling process but as

an initial step this study explored the impact of the alignment process on a small sample of teachers with diverse exposure to mathematics and test development.

Methods for Making Contact

Through my work in the test development process I had collaborated with ABE teachers on a number of different projects. The two test development participants were extensively involved in a revision to the Math ABE Standards process (Martone, Goodridge, Moses, & Titzel, 2004) and in the item review process. They brought a solid understanding of the methods behind the test development process. Through my work as a representative to the ACLS Math Professional Development Initiative I contacted the TIAN representatives and invited two teachers from this group to continue their exploration of the standards through this alignment study. Finally, through my work with ACLS I obtained recommendations of teachers who have expressed an interest in math but did not join the previously mentioned activities. I expected this last group to have less knowledge about assessments and standards but to possibly be more representative of typical Math ABE teachers. I ensured that at least two participants were from this category.

My initial contact was through an emailed description of the project, the time requirements, and compensation information. Once the participants were selected I followed up with a more detailed description of the project, the timeline, and the participant requirements. I mailed two copies of an informed consent form to each participant so they could review it prior to initiation of the study. A sample of the informed consent form is in Appendix B.

Data Gathering Methods

A mixed method study allows for a combination of methods to provide data as a form of triangulation. The results from each method help to support the generalizability of the results. This study used an analysis of the degree of alignment through item and objective coding using the Webb methodology, discussions with the participants throughout the alignment process and a videotaped focus group discussion to learn about the degree of alignment and the influence the alignment process had on the participants. This study occurred in three phases with the results from one phase connected to the results of another phase. In this way the study had "convergence, corroboration, and correspondence of results between the different methods" (Johnson, 2006). Additionally, the use of probing questions throughout and the concluding focus group discussion allowed for "elaboration, enhancement, illustration, clarification of results from one method with the other method"(Johnson, 2006). The data in this study were gathered concurrently and given equal weight in the analysis. Using the mixed method approach enabled the results from one phase of the study to guide and elaborate on another phase.

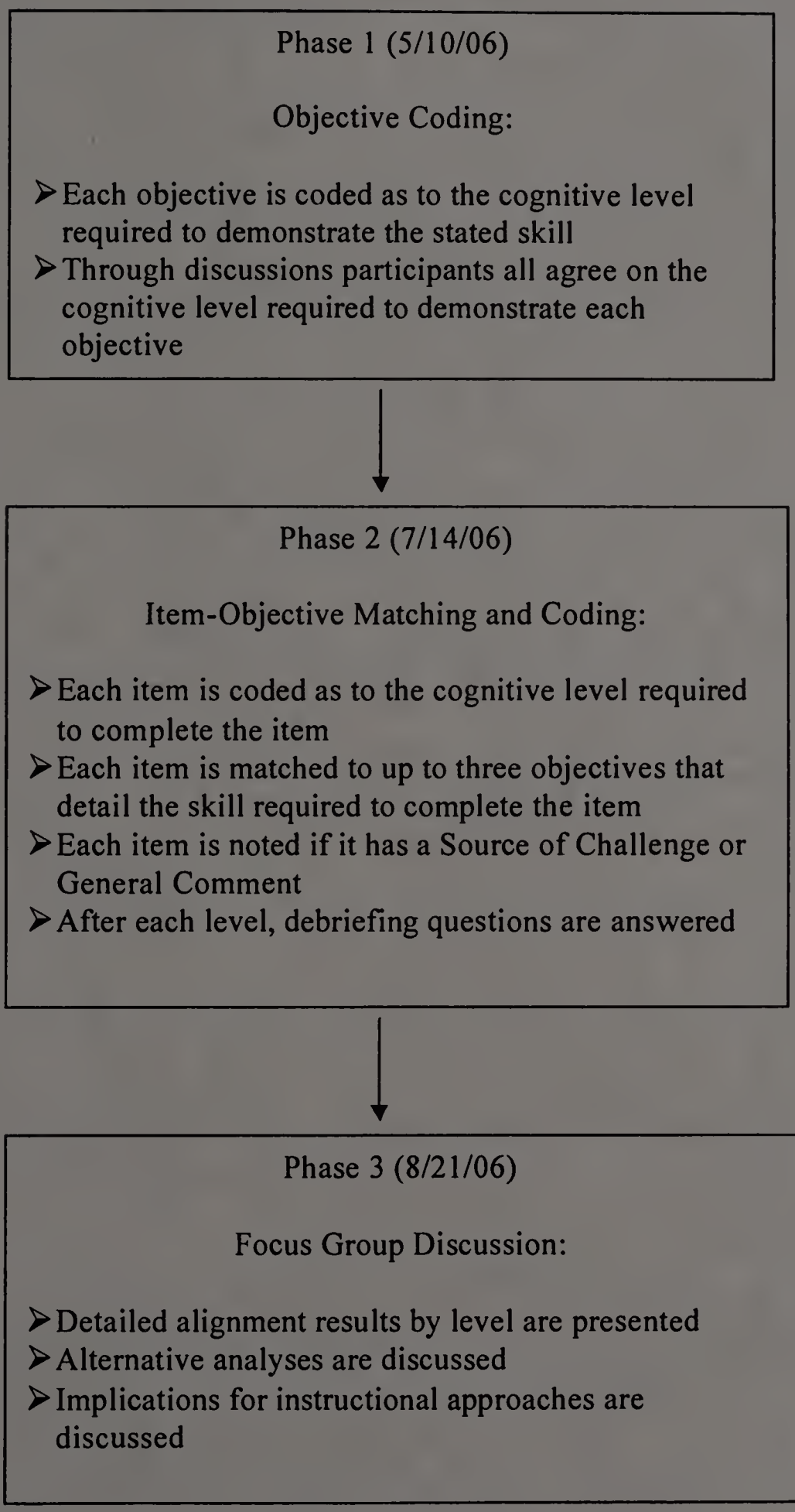
This study involved three phases, which provided the data for the mixed method approach to data collection and analysis. The first phase entailed a detailed review of the Math ABE standards where each objective was coded based on one of the three cognitive levels required to demonstrate that skill (Knowledge & Comprehension, Application, or Analysis, Synthesis & Evaluation). The codes for each objective were then discussed among the group until a consensus cognitive level was determined for each objective. Throughout this phase I noted observations about the participants'

interactions and discussions. Following this phase the participants were asked probing questions to understand how this task influenced their understanding of the Math ABE Standards.

The second phase required the participants to review each item within each test level, match the item to one of the three cognitive levels (Knowledge & Comprehension, Application, or Analysis, Synthesis & Evaluation), and then match the item to up to three objectives. The participants also noted any source of challenge concerns or general comments they had for each item. After coding each level, the participants were asked to respond in writing to three debriefing open response survey questions about their view of the MAPT for Math and to answer a Likert-type survey question summarizing the degree of alignment between the MAPT for Math and the Math ABE standards. Throughout this second phase of data collection observations were again noted about the participants' interactions and discussions although this phase was a more independent activity than phase one.

The third phase was a videotaped focus group discussion about the alignment results and the overall alignment process. The results for each data collection phase and each test level were presented. I also discussed with the participants other possible approaches to data analyses. The focus group discussion also included questions about how the alignment process influenced the participants' approach to instruction. The process of data collection for each of these phases will be discussed in the next section and is summarized in Figure 4 as both a flow chart and a table showing the steps and the related results.

Figure 4 – Outline of Data Collection Phases



Summary of Data Collection Methods and Results

Phase	Data Collection Method	Results
Phase 1	Code Math ABE Standard objectives for a cognitive level	Consensus cognitive level rating for each objective
	Probing questions and observations	Initial participants understanding of Math ABE Standards and a better understanding of how the process worked
Phase 2	Alignment coding results	Cognitive level items and item/objective matches
	Probing questions and debriefing questions	Increased participant understanding of Math ABE Standards and MAPT for Math
Phase 3	Focus group discussion of the results and the overall process	Videotape of the results and questions asked throughout and the teachers' interactions, comments, and questions

Phase 1 – Objective Coding

The first phase began with a focus on the Math ABE standards. The results for this phase required that each objective within the Math ABE standards have a consensus cognitive level required to demonstrate that skill. To accomplish this task each participant first independently reviewed the objectives at each level and coded what depth of knowledge is required to accomplish that skill.

There are four strands within each learning level. These strands are Number Sense (N), Patterns, Functions, & Algebra (P), Statistics & Probability (S), and Geometry & Measurement (G). Within each of these standards are objectives that detail

what skills are related to that strand for each level. The information in Table 3 details the number of objectives per strand per level.

Table 3 – Objectives within Each Strand within Each Level

Learning Level	Strand	Objectives
2	N	19
	P	11
	S	15
	G	15
	Total	60
3	N	25
	P	14
	S	30
	G	19
	Total	88
4	N	29
	P	17
	S	30
	G	23
	Total	99
5	N	17
	P	11
	S	24
	G	14
	Total	66
Grand Total		313

Phase 1 began with a training to review the cognitive level terms. The cognitive level terms are Knowledge and Comprehension (KC), Application (App), and finally, Analysis, Synthesis and Evaluation (ASE). These cognitive levels represented a collapsed version of Bloom' s(1956) taxonomy and are the classifications that were used to develop the MAPT for Math. The participants were provided with definitions for each cognitive level as shown in Appendix C. After reviewing the definitions the participants coded 10 objectives as to the cognitive level required for each one. The results of these 10 objectives were discussed among the whole group to ensure there

was a common understanding of the cognitive levels and how they related to the objectives. The participants then wanted to code a standard independently (Level 3 Number Sense) and then revisit it as a group. Any objectives that were not unanimous were discussed until consensus was reached. Then the participants coded the other three strands within level 3 and again any non-unanimous objectives were discussed. The participants then coded the objectives within the other learning levels and discussed the results where necessary. This coding process helped the participants to become deeply familiar with all of the objectives at each learning level. The discussion and debate required to reach consensus regarding the depth of knowledge categories also helped the participants to have a deeper understanding of the cognitive level terms and how they were operationalized in the objectives.

Phase 2 – Item-Objective Matching and Coding

The item-objective matching and coding process involved a number of steps for the participant. First, each item was coded to one of the three cognitive levels used in the objective coding process. Second, each item was coded to up to three objectives. Third, the participants noted if an item had a source of challenge issue or a general comment. Fourth, the participants completed survey debriefing questions after each level. Each of these steps was reviewed with the participants in the training prior to the implementation of phase 2. The training and coding process details will now be discussed.

Alignment training. There were four steps to the training process for the item-objective matching and coding. First, the training began with a review of the cognitive levels. Because the objective coding process was two months earlier, the participants

needed to review the cognitive level definitions. This review entailed discussing the cognitive level descriptions, presenting examples from the objective coding process, and discussing some of the lessons learned from the objective coding process. Then participants coded every sixth item from the level 3 to a cognitive level. The results of this coding were then discussed among the group.

Second, the participants were trained as to how to match items to objectives. It was important that the participants understood how items must be matched to the objective that most fully represents what the item is testing. Items could be matched to up to three objectives (one primary and two secondary) if the participant truly thought the item was fully measuring more than one objective. Participants also noted if they thought the item could not be matched to any objective. To facilitate the item-objective matching process a new presentation format of the standards was used. The original way the standards document was produced was as a list of standards and objectives (http://www.doe.mass.edu/acls/frameworks). Unfortunately, in this presentation format, it would have been very difficult for the participants to find the objective to which an item matches, especially if there are similar objectives that are not listed near each other in the document.

Due to the length of the Math ABE standards, the matching process was facilitated by a table view of the standards and objectives based on common topics. Because of the quantity of objectives, the table was a more systematic way to view the complete objectives across levels. Additionally, given the adaptive nature of the test, each test level might have items from another test level so this view facilitated the identification process between levels. The table view groups objectives by topical areas

(these are the rows) and then shows how objectives related to that topic change across the levels (these are the columns). The complete table view of the standards can be found at http://www.doe.mass.edu/acls/assessment. A sample table view is in Table 4.

Table 4 – Sample Table View of the Math ABE Standards

Standard	Topic	2	3	4	5
Number Sense	Fractions	2N-1.3 Read, write, and compare halves and quarters of quantities. 2N-2.5 Know halves of even numbers up to 100.	3N-1.2 Read, write and compare common fractions (e.g. thirds, halves, quarters).	4N-1.3 Read, write, order and compare fractions and mixed numbers. 4N-1.12 Recognize and use equivalent forms of common fractions (e.g.1/2 = 5/10)	5N-1.2 Read, write, order and compare fractions and mixed numbers. 5N-3.3 Add, subtract, multiply and divide using fractions and mixed numbers.
Number Sense	Fractions - manipulating		3N-2.1 Demonstrate an understanding that multiplying a whole number by a unit fraction is the same as dividing the whole number by that fraction' s denominator. 3N-3.6 Find common parts of whole number quantities or measurements (e.g. ¾ of 12, 2/3 of 15).	4N-3.3 Evaluate one number as a fraction of another. 4N-3.4 Use common fractions to add, subtract, multiply and divide amounts or quantities.	5N-2.1 Demonstrate an understanding of the effects of each operation with fractions.
Number Sense	Decimals		3N-1.4 Read, write and compare decimals up to two decimal places in practical contexts (such as money in decimal notation, e.g. $10.35). 3N-3.4 Carry out basic calculations with money.	4N-1.4 Read, write, order and compare decimals up to three decimal places. 4N-3.2 Add, subtract, multiply and divide decimals up to three places.	5N-1.3 Read, write, order and compare decimal numbers of any size. 5N-3.1 Add, subtract, multiply and divide decimals of any size.

Once a participant identified the topic of the item, he/she could then go to that area of the table and focus on the level that is associated with the test. Then if an objective for a match was not found, he/she could easily look at the adjacent levels to see if there is a possible match at a higher or lower level. For example, one topic is fractions and then the objectives across that row show how this skill is assessed at the different learning levels. While this layout is different from how the standards were originally produced and distributed to the teachers, the Math ABE standards were new enough so teachers were not too attached to the original list format. ACLS has also included this table view of the standards as a reference on their website as noted above.

In the training, the participants learned about the table layout of the standards and practiced using it to identify item/objective matches. Every 6th item from the level 3 tests was matched to up to three objectives. Then the participants discussed their matches and their reasoning for each match. Through this discussion the participants also referenced the table view of the standards and how they focused on the topic being tested to then find the objective that is being tested.

Third, the participants learned about examples of items that presented source of challenge issues. Examples of these items were presented to the participants and the group discussed other possibilities. Unfair sources of challenge are situations where students who know the item might still get the item wrong or students who do not know the item might still get it right based upon the way the item is presented. For example, if a math item requires excessive and unnecessary reading this might unfairly challenge students so they might not be able to show their true understanding of the concept being tested. Another example would be a math item that asks for the perimeter of a three by

six rectangle might be a source of challenge because students could confuse area and perimeter and still get the item correct. The participants were also told they could make any general comments about item or an item/objective match.

The final step of the training was a review of the debriefing questions the participants were asked at the completion of each level. A sample of the debriefing questions is included in Appendix D. There were three open-response debriefing questions and each was discussed with the participants during the training. These questions were: 1) For each standard, did the items cover the most important topics you expected? If not, what topics were not assessed that should have been? 2) For each standard, did the items cover the most important cognitive levels you expected? If not, what cognitive level was not assessed? 3) Was there any content you expected to be assessed, but you found no items assessing that content? What was that content? Then the participants were asked to state their general opinion of the alignment between the standards and the assessment for that test level. Their opinion was captured through a five point rating scale ranging from perfect alignment to not aligned in any way. Then the participants could note any general comments they had. The debriefing questions were reviewed and discussed with the participants before the item-objective matching and coding process began.

The training process provided the participants with all of the information they needed to complete the coding process. By the end of the training the participants learned about how to code items to a cognitive level, how to use a new presentation of the standards and objectives, how to match items to objectives, what might cause an

item to have a source of challenge issue, and how to complete the debriefing questions. At this point the participants were ready to begin the alignment coding process.

Coding process. There were four steps to the coding process. These steps required the participants to code each item as to the cognitive level required to complete that item, match the item to up to three objectives, note if the item had a source of challenge, and note any general comments for that item. A sample coding form is in Appendix E.

The six participants each reviewed the same complete 40 item test at each of the four learning levels to result in a total of 160 items. The participants completed the rating process for each level before proceeding to the next testing level. Unlike the objective coding process in phase 1, the participants did not discuss their results or reach a consensus for any of the item-objective matching. The participants coded each item independently and the results were averaged across participants.

The first step of the item-objective matching process began as the participants assigned one of the three cognitive levels to each item. This process was similar to the rating of the objectives but the results of this coding process were not discussed and a consensus was not needed. The second step of the item-objective matching process was for the participants to match each item to a primary objective and up to two secondary objectives. The third step of the item-objective matching process was for the participants to note any items that had a source of challenge issue. Participants could also note in a separate field any general comments they had about the item, the objective, or their thinking. The fourth and final step of the item-objective matching

and coding process was for the participants to complete the debriefing questions after each level.

The participants completed test levels 3 and 4 during the one day meeting on July 14th, 2006. On this day all coding was done on paper and only paper versions of the Math ABE standards were available. Due to time limitations, the participants took home levels 2 and 5 to complete the item-objective matching and coding process. While coding at home, the participants had access to an electronic version of the table view and list view of the Math ABE standards. Prior to adjourning the meeting, the participants discussed how the electronic version of the Math ABE Standards in table form could be searched using the find feature in Word to facilitate the matching process. The participants discussed examples of how key words could be used to hone in on the objectives that related to what an item was asking. All items, coding sheets, and debriefing questions were returned within a week of this meeting.

Phase 3 – Focus Group Discussion

In the final phase of this study the participants learned about the results of the alignment study and discussed the results and the overall alignment experience. This meeting took place about three weeks after the item-objective matching process to allow time for data analysis but to ensure that the experience was still fresh in the participants' minds. The presentation of the alignment results and the ensuing discussion was videotaped to allow for additional review of the data after the group concludes.

While listening to the results, the participants were encouraged to discuss the results and how the results related to their view of the Math ABE Standards, the MAPT for Math, and their approach to instruction. Specific probing questions were written

that built on the findings from the observations throughout the alignment process, the earlier discussions among the participants, and the results for the different alignment dimensions. After the results were presented, the participants were asked some summary questions about the alignment process and results as a whole and any instructional effects the process and results might have. These questions included: What did you learn from this process that will help you in your classroom? Were you surprised about any of the results? What was interesting about this process? What was challenging about this process? What would you change about this process? The meeting concluded with a discussion of some preliminary concerns with the Webb dimensions and some alternative ways to analyze the data.

Data Analysis

Two different types of data analysis occurred in this mixed methods study. First, to answer the first research question, the degree of alignment between the MAPT for Math and the Math ABE Standards was analyzed across the four dimensions using average ratings and cutoff criteria determined by Webb (Webb, Alt, Ely, & Vesperman, 2005). The source of challenge, general comments and debriefing questions were also analyzed as a means to examine the degree of alignment. Second, to answer the second research question, the observations and discussions from throughout the alignment process and the focus group videotaped discussion were analyzed using open, axial coding techniques. Each of these analyses will now be discussed.

Alignment Criteria

Four alignment dimensions from Webb et al. (2005) were calculated using the results of the alignment study. These dimensions included: categorical concurrence,

depth-of-knowledge consistency, range of knowledge correspondence, balance of representation. The analysis for each dimension is detailed below.

An important aspect of the Webb methodology is the term "hit". A hit is any item/objective match. Given that participants could match an item to up to three objectives, each item could potentially have three hits. The hits do not need to be within the same standard. So an item could be matched to an objective within Number Sense and also an objective within Statistics and Probability. Understanding this terminology is an important foundation for the analyses that follow.

Categorical concurrence. Categorical concurrence compared the similarity of the expectations for student learning, as expressed through the content categories in the standards, to the assessments. The total number of item/objective matches, hits, within a standard was averaged across all participants to determine the average number of items per standard. To have alignment relative to this dimension, an assessment must have had at least six items measuring a standard. Using this approach, if there were four standards, an assessment needed at least 24 items with six items per standard to determine there was alignment relative to categorical concurrence. Webb et al. (2005) detailed the rationale for the six item criteria for categorical concurrence. They stated,

> Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63. This indicates that about 63% of the group would be consistently classified as masters or nonmasters if two equivalent test administrations were employed. The agreement coefficient would increase if the cutoff score is increased to one standard deviation from the mean to .77 and, with a cutoff score of 1.5 standard deviations from the mean, to .88. Usually, states do not report student results by standards, or require students to achieve a specified cutoff score on subscales related to a standard. If a state did do this, then the state would seek a higher agreement coefficient than .63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a standard and as a basis for

> making some decisions about students' knowledge of that standard. If the mean for six items is 3 and one standard deviation is one item, then a cutoff score set at 4 would produce an agreement coefficient of .77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

The meaningfulness of the six item cutoff will be discussed in greater depth in the discussion section.

Depth-of-knowledge consistency. Depth-of-knowledge consistency compared the level of cognitive demand expressed in the specific objectives within each standard to the cognitive demand in each item that is matched to that objective. The main criterion here was that what was tested should be at or above the same cognitive level as what is expected to be taught. This dimension was calculated for each standard for each participant and then the results for percentage of assessed objectives in each category are averaged across the participants.

This could be a confusing calculation so an example is provided to illustrate the calculation. The hits for Judy for standard 4 Number Sense are shown in Table 5. The table shows the Standard, the objectives Judy matched items to, the number of items that were under (UN), at (AT), and above (AB) the objective to which it was matched, and then the percentage distribution of those hits for that specific objective in terms of percent under (UN), percent at (AT), and percent above (AB). Objective 4N-2.1 illustrates an important point about percentage. Two of the items matched to this objective are at the cognitive level of the objective and one item was above the cognitive level of the objective. Therefore, for this objective 67% of the items were at and 33% of the items were above. The average percentage in each category was then

calculated using the total number of objectives within that standard (for the example below that is 33 objectives).

Table 5 – Item/Objective Matches for Judy for Level 4 Number Sense

Standard	Objective	# UN	# AT	# AB	% UN	% AT	% AB
4N	4N-1.1	0	1	0	0.00	1.00	0.00
4N	4N-2.1	0	2	1	0.00	0.67	0.33
4N	4N-2.2	0	0	1	0.00	0.00	1.00
4N	4N-2.4	0	2	0	0.00	1.00	0.00
4N	4N-2.5	0	1	0	0.00	1.00	0.00
4N	4N-3.4	0	0	1	0.00	0.00	1.00
4N	4N-3.5	0	0	1	0.00	0.00	1.00
4N	4N-3.6	0	1	0	0.00	1.00	0.00
Average					0.00	0.14	0.10

The analysis was repeated for each participant for each standard and then the average across all participants was calculated as shown in Table 6. The total average percent under, at, and above across all the participants was then calculated. The final step was to determine the percent under, at, and above as a percentage of that total.

Table 6 – Average Hit Distribution by Standard by Participant

Standard	% UN	% AT	% AB	Participant
4N	0.03	0.12	0.12	Beth
4N	0.00	0.14	0.10	Judy
4N	0.00	0.09	0.12	Len
4N	0.06	0.15	0.09	Mary
4N	0.06	0.09	0.18	Melissa
4N	0.06	0.20	0.08	Sabrina
Average	0.04	0.13	0.12	
Sum				0.28
Percentage	0.13	0.47	0.41	

The sum of the percentage at and above across all participants must be greater than or equal to 50% to meet the requirements for acceptable depth-of-knowledge consistency. This example would meet that requirement (0.47+0.41=0.88 of the assessed objectives are assessed by items that are at or above the cognitive level of that objective). Fifty percent was based on the assumption that most cutoff points require students to answer more than half the items to pass. The main concern with this aspect of alignment was that assessment items should not be targeting cognitive skills that were below those required by the objectives.

Range of knowledge correspondence. Range of knowledge correspondence analyzed the breadth of the standards as compared to the breadth of an assessment. This aspect of alignment looked at the number of objectives within a standard measured by at least one assessment item. To have sufficient alignment relative to range of knowledge,

at least 50% of the objectives within a standard needed to be measured by at least one assessment item. This criterion assumed that students should be tested on at least half of the domain of knowledge. This part of the alignment process also assumed all of the objectives have equal weighting and all of the objectives accurately cover the skills needed to complete that standard.

Balance of representation. Balance of representation focused on how evenly assessment items were distributed across objectives within a standard to represent the breadth and depth of the standards. This aspect of alignment focused on the objectives that were assessed by an item and then examined the proportion of objectives measured compared to the number of items. The goal was to measure every assessed objective with at least two items. Specifically the calculation for the balance index for each standards is: $1-(\sum_{k=1}^{0}|1/(O)-I_k/(H)|)/2$, where O=Total number of objectives hit for the standard; $I_{(k)}$ = Number of items corresponding to objective (k); and H = Total number of items hit for the standard (Roach et al., 2005). Table 7 shows a sample of this index for one standard for one participant.

Table 7 – Sample Balance of Representation Calculation

Standard: Level 4 Number Sense
Participant 1

Assessed objectives	Items per Obj ($I_{(k)}$)	Absolute value of (1/O)-($I_{(k)}$/H))
4N-1.1	1	0.03
4N-2.1	3	0.15
4N-2.2	1	0.03
4N-2.4	2	0.06
4N-2.5	1	0.03
4N-3.4	1	0.03
4N-3.5	1	0.03
4N-3.6	1	0.03
	Sum of Abs Value	0.41
	Divide by 2	0.20
	Subtract from 1	**0.80** **Balance Index**

# Objs (O)	Hits (H)
8	11

The overall balance index for this standard would then be the average balance index across the six participants. If the proportion approached zero that signified many items were assessed by only a small number of objectives within a standard. If the proportion approached one that signified that the assessed objectives were matched to an equal number of items. A balance index of 0.7 or higher represented a balanced standard with items fairly evenly distributed among the objectives that were measured. Index values of 0.6 to 0.7 represented a weakly met balance of representation criterion.

Source of challenge. Source of challenge was met if the primary difficulty of the assessment item was significantly related to students' knowledge and skill in the content area as represented in the standards. This information was recorded as notes next to each item as it was matched to an objective. Then the proportion of items having source of challenge issues was noted. Webb does not recommend a specific cutoff point for this dimension as to what signifies too many items with a source of

challenge concern. On the individual item level these items were reviewed and modified as needed for future test administrations.

General comments. Throughout the item-objective matching process the participants made general comments. The general comments allowed the participants to capture their thinking in a way that wasn' t possible with limited interaction among the participants during the coding process. The general comments were reviewed as a whole and then coded using open, axial coding.

Debriefing questions. After each level was rated, the participants completed debriefing questions in writing. The results to these questions were combined and analyzed to look for the commonality across the findings. Given the limited number of participants and the list form of the answers, the results to these questions were incorporated in full. The participants also ranked the degree of alignment between each assessment level and the matching standards on a scale of one to five. The percentage distribution for each assessment level was analyzed.

Alignment to Test Specifications

The above Webb methodology application to analysis focused on the alignment of the assessment to the standards but does not actually examine how well the assessment accomplishes what it was designed to do. For example, Webb' s categorical concurrence requires six items per standard. In the test design process some standards are purposefully weighted differently. Therefore, an additional analysis compared the results of the Webb criteria to the specifications detailed in the MAPT for Math test specifications which were shown in Table 2. The test specification table is the

document that connects the instruction, the curriculum, and the assessment as it sets out the relative emphasis for each strand and cognitive level.

In the Webb methodology, acceptable levels of categorical concurrence and depth-of-knowledge consistency are predetermined. The results for these dimensions were compared to the original test specifications and presented to the participants in the focus group discussion. Additionally, the Webb approach did not account for level of participant agreement beyond the reliability calculations. By averaging hits across standards, the Webb methodology attempted to decrease the impact any aberrant participant coding might inflict.

Webb has recently agreed that it is important to assess the degree of agreement among participants as to how items are coded (Webb et al., 2006). This study found that using a minimum threshold of participant agreement, rather than including all the matches and averaging across participants, led to different views of alignment in terms of categorical concurrence and range of knowledge. The study found that requiring agreement at the objective level may be too stringent for categorical concurrence, but is necessary for determining adequate range of knowledge correspondence. In terms of mapping back to the original test specifications, agreement at the strand level is most important and, as this study notes, can have considerable impact on what items are included in the alignment analysis. For this study, an additional analysis looked at how categorical concurrence and depth-of-knowledge consistency would be met if a level of participant agreement was required and the results were compared to the requirements set forth in the test specification table. To determine alignment to the test

specifications, only items that had at least four out of six participants' agreements about the strand or cognitive level classification were included.

Discussion and Observation Analysis

The discussions and observations throughout the alignment process, as well as, the videotape of the focus group discussion were analyzed to determine the results for the second research question. After each phase of the alignment process the participants informally and through email shared their thoughts about the activities. I also recorded observations about participants' interactions throughout the phases of the alignment process. The discussions and observations were transcribed immediately following each meeting. Then notes were made about themes or ideas to explore in future meetings with the participants. The videotape of the focus group discussion was transcribed to capture the participants' comments, questions, and interactions.

The transcriptions were then coded using open, axial coding to inductively develop categories and explore themes (Creswell, 1998). Some of the themes and categories started to develop through the earlier discussions and observations and were explored, developed, and modified while new categories were also created. The open coding was an initial review of the data and a beginning step in assigning categories to the findings. Through this process I moved away from the specific questions that I asked to look across the data for common categories and themes. With the axial coding I made new connections between the categories to form ideas about how the concepts work together to develop thematic findings (Creswell, 1998). The result of this analysis was a full integration of the informal discussions, observations, and formal focus group

data to illustrate the impact the alignment process had on the participants' view of the MAPT for Math, the Math ABE Standards, and their instruction.

Trustworthiness

There were multiple approaches to ensure the trustworthiness of the findings. Trustworthiness relates to reliability and was examined through both statistical analyses and methods more common in the qualitative field. For the Webb criteria alignment analysis I used two forms of reliability to analyze the reliability of the participants' ratings. To limit my researcher bias and ensure the credibility of my qualitative findings I used three methods recommended by Rossman and Rallis (2003): (a) triangulation, (b) participant validation, and (c) a peer debriefer.

Reliability Measures

Webb recommends two forms of analysis to examine the reliability of the participants' ratings. For the item cognitive level classification I measured the participant intraclass reliability to examine the correlation among participants in assigning cognitive categories to the items (Webb, Alt, Ely, & Vesperman, 2005). Webb et al. (2005) used the Shrout and Fleiss (1979) method for intraclass correlations as stated below,

$$ICC = \frac{\sigma^2(i)}{\sigma^2(i) + \sigma^2(r)}$$

> Here, $\sigma^2(i)$ is the variance in the data between the assessment items, and $\sigma^2(r)$ is the variance in the data between the participants. In other words, the statistic measures the percent of variance in the data due to the differences between the items rather than the differences between the participants. An intraclass correlation value of, say, 0.7, means that 70% of the variance in the data can be explained by differences between the items while the other 30% is due to differences between the participants.

Values greater than 0.8 represented good intraclass correlations and values between 0.7 and 0.8 represented acceptable intraclass correlations.

The reliability of the item-objective matches was calculated using average pairwise comparisons. In Webb' s methodology he recommends that items are matched to one primary objective and up to two secondary objectives and that was replicated in this study. Webb et al. (2005) provided the steps to calculate the pairwise comparison as follows:

1) For a pair of participants, find the number of objectives the two participants agreed on and divide that number by the total possible number of matches. For example, participant A coded an item to 2N-1.1 and 2N-1.3 and participant B coded the item to only 2N-1.1. The agreement for these two participants for this item is 1/2. This is the agreement between two participants for a single item.
2) This is repeated for all possible participant pairings.
3) Sum all of the participant agreement values for this item. Divide this sum by the total number of pairs of participants. This is the pairwise agreement value for a single assessment item.
4) Average all of the pairwise agreement values across all of the items to get the pairwise agreement objective for the whole alignment study.

An average pairwise agreement measure of 0.6 and above was deemed good. And values of 0.5 to 0.6 were seen as acceptable.

Triangulation

As a means to understand how teachers' participation in an alignment review process influenced their view of the standards, the assessment, and their approach to instruction I studied their involvement and reaction to the alignment process at multiple points. I gathered observations throughout the alignment process. I also asked probing questions at different points to explore the participants' thinking. The debriefing questions gathered open-responses to further explore these questions. And finally, I videotaped the participants' involvement in a presentation of the results and the focus

group discussion about the alignment process. The different sources of qualitative data helped to explore the consistency of viewpoints or provide examples of how and why viewpoints changed through the alignment study. The triangulation of the data collection points bolstered the trustworthiness of my findings and interpretations.

Participant Validation

Participant validation, also known as member checks, is the process of sharing the interpretations of the findings with participants (Rossman & Rallis, 2003). After I analyzed the data and developed the results I presented my results and conclusions to my participants. This helped ensure that I did not misrepresent my participants' words or actions. The member checks also provided a valuable opportunity to relate specific examples to the broader context of the participants' experiences. If Ihad any questions about why a situation or comment occurred I used the member checks as a means to learn more about what might not have been apparent in the data I had available. The participant validation process augmented the credibility of the findings from this study.

Peer Debriefer

Throughout this research study I consulted with a peer to review my methodology and share ideas about my analysis. This peer helped with videotaping process and was therefore present during the data collection. He served as an "intellectual watchdog"(Rossman & Rallis, 2003, p. 69) throughout the process and this lent credibility to the research findings.

Researcher Bias

Since part of this study involves subjective interpretation of data collected by the researcher, it is important to understand my personal background with this research

question and any bias I might have brought to the data collection and analysis process. My research interest developed from my experience with assessment, standards, and professional development around instructional change as a fourth grade teacher in New York City and through my work with pre-service teachers. I have worked with the ABE population since 2004 on the development and implementation of standardized, computer-based assessments for Math and Reading. In this role I interacted with teachers in a standards revision project, on item writing, and on item review. Through the teachers' involvement in these steps I saw how they began to take ownership of the assessment and have a voice in how it developed.

Personally I wanted to see teachers' involvement in the alignment process have a positive impact on their view of assessments, standards, and their approach to instruction. I believe that having teachers "at the table" will help them to have their voices heard as a means to influence future changes to the assessment and the standards as well as learn what additional types of professional development might be needed to support instructional change. I used discussions with my advisors and my peer debriefer to help me to see when my biases might have impeded my analysis and interpretation of the data.

I think I was effective at helping my participants to feel comfortable as I observed them in the alignment process and interviewed them in the focus group setting. I was conscious not to be overly supportive or critical of anything I saw throughout the experience while I also worked to represent the participants' voices accurately. It was important to me to provide all the participants with results and discussion points so they could ensure I adequately represented their experience. This

is a population I have enjoyed working with in the past and hope to continue working with in the future. I wanted to ensure I did not do anything to hurt any of the participants or jeopardize the relationships we have built.

CHAPTER 4

RESULTS

Introduction

This study explored two related research questions. The first research question asked how well the MAPT for Math aligned with the Math ABE standards. This question was answered through an application of the Webb methodology by six ABE math teachers. The second research question explored the impact of the teachers' participation in the alignment process on their views of the standards, the assessment, and their instruction. This question was answered with discussions and observations throughout the alignment process, formal debriefing questions, and a focus group discussion where the results from research question one were presented. The results for each research question will be presented separately.

Results for First Research Question: To what extent is the MAPT for Math aligned to the Math ABE standards?

The results for the first research question are presented in five main sections. First, the results from the objective coding process are presented. Second, overall alignment and reliability results are detailed. Third, the results for each of the four levels of the assessment across the four alignment dimensions are presented. Fourth, the last section of results regarding how well aligned the MAPT for Math is to the Math ABE standards looked at the results from the qualitative components of this analysis. These components included the source of challenge comments, the general comments, and the debriefing questions for each level. Fifth, the results of the Webb methodology findings compared to the test specification table are presented.

Objective Coding Results

The participants' consensus ratings regarding the cognitive level for each objective within the Math ABE standards are presented in Table 8. This table illustrates the percentage of objectives that require Knowledge and Comprehension (KC), Application (App), and Analysis, Synthesis, and Evaluation (ASE) expectations within each learning level.

Table 8 – Distribution of Objectives' Cognitive Levels Across Learning Levels

	Learning Levels			
	2	3	4	5
# of Objectives	60	88	99	66
Cog. Level				
KC	50%	44%	54%	30%
App	42%	40%	33%	38%
ASE	8%	16%	13%	32%

From the table the participants saw a heavy concentration of Knowledge and Comprehension type objectives at levels 2 through 4 with less than 20% of the objectives at these levels requiring Analysis, Synthesis, and Evaluation type skills. Level 5 had objectives that required a more even balance across the spectrum of cognitive levels. The participants thought it was reasonable that level 2 had few Analysis, Synthesis, and Evaluation objectives (8%) but then this increased in level 3 (16%). The participants noted that the percentage of Analysis, Synthesis, and Evaluation then dropped at level 4 (13%) while Knowledge and Comprehension rose from 44% to 54% between levels 3 and 4.

Beth, who was deeply involved in the framework creation, was not surprised with the change between levels 3 and 4 when the type and difficulty of the content was considered. She said, "Maybe [the increase in Knowledge and Comprehension] is

because we are presenting new material at level 4" so these ideas should start at a more basic level. Then the topics are extended in level 5 where ASE expectations for the objectives become more prevalent (32%). Mary also noted, " [Knowledge and comprehension expectations] make sense where it is at a beginning level [of a topic]. [Students] need to have the knowledge before you can get to that ASE." Participants looked at the results within the levels of the Math ABE standards and considered what was happening in their classroom to determine if the results seemed reasonable to them.

Table 9 illustrates the percentage of objectives for each level in terms of both strand and then cognitive level within that strand. For example, of the 60 objectives for level 2, 32% are for Number sense and of those 18% are KC, 13% are App, and zero objectives are ASE.

Table 9 – Distribution of Objectives' Cognitive Levels Across Strands/Learning Levels

		Levels			
		2	3	4	5
# of Objectives		60	88	99	66
Strand	Cog. Level				
N	KC	18%	16%	19%	15%
	App	13%	13%	10%	11%
	ASE	0%	0%	0%	0%
Total		32%	28%	29%	26%
P	KC	10%	8%	10%	5%
	App	8%	5%	3%	8%
	ASE	0%	3%	4%	5%
Total		18%	16%	17%	17%
S	KC	8%	13%	14%	8%
	App	12%	10%	9%	6%
	ASE	5%	11%	7%	23%
Total		25%	34%	30%	36%
G	KC	13%	8%	10%	3%
	App	8%	13%	11%	14%
	ASE	3%	1%	2%	5%
Total		25%	22%	23%	21%
Grand Total		100%	100%	100%	100%

Two of the participants noticed a greater proportion of Analysis, Synthesis, and Evaluation objectives at the higher levels in the Statistics and Probability (7% at level 4 and 23% at level 5). Beth shared how the strands within the Math ABE standards were written by different teams of people so the difference in the writers' understanding of student expectations was operationalized in terms of cognitive expectations within the different strands. Len said, " The frameworks also revealed something I hadn' t really suspected. How uneven the frameworks are...I kind of thought they were a whole thing. I didn' t realize there were different sections that were worked on by different people at different times. And that the language was going to be so different from section to section." Looking at the objectives through the lens of cognitive expectations helped the participants to see differences among the strands and think about how their instruction relates to the standards.

The participants noticed that very few of the objectives at each level had ASE cognitive expectations. The participants did have suggestions for modifications based on their approach to instruction and different student populations. Beth wanted to see more Analysis, Synthesis, and Evaluation at the lower levels to help extend what students are learning. She witnessed this level of thinking occurring in the classroom and thought the expectation should be expressed in the standards for all teachers.

Sabrina agreed that having more ASE at all levels could be helpful to set the expectations for teachers. She stated,

> " My overall opinion of the complexities of the objectives is that while level 5 objectives challenge the learners to use more complex thinking skills, there might be a place for more of those higher order skills at the 2,3 and 4th levels. I see a wonderful trend in the classroom that encourages learners at all levels to investigate and discover mathematical precepts. But, the objectives seem to imply that if the learner can "compute" then we are satisfied that he's

> achieved success at the lower levels. I know this is not the "objective" of all those wonderful math teachers who worked on those frameworks."

While the cognitive level expectations in the frameworks help teachers to think about how to teach the skills, the differences among the levels and strands shows that these distinctions may need to be revisited.

Alignment Dimension Results

The alignment results using the Webb methodology criteria for each dimension for each standard are shown in Table 10. Standards that do not meet the criteria for that dimension are shaded in dark grey and weakly met dimensions are in light gray.

Table 10 – Summary View of Alignment Based on the Four Webb Dimensions

Standard (Level – Strand)	Depth-of-Knowledge Consistency	Balance of Representation	Categorical Concurrence	Range of Knowledge
2 – Number Sense	Yes	Yes	Yes	Weak
2 – Patterns, Relations, and Algebra	Yes	Yes	No	No
2 – Statistics and Probability	Yes	Yes	Yes	No
2 – Geometry and Measurement	Yes	Yes	Yes	No
3 – Number Sense	Yes	Yes	Yes	No
3 – Patterns, Relations, and Algebra	Yes	Yes	No	No
3 – Statistics and Probability	Yes	Yes	Yes	No
3 – Geometry and Measurement	Yes	Yes	Yes	No
4 – Number Sense	Yes	Yes	Yes	No
4 – Patterns, Relations, and Algebra	Yes	Yes	Yes	Weak
4 – Statistics and Probability	Yes	Yes	Yes	No
4 – Geometry and Measurement	Yes	Yes	Yes	No
5 – Number Sense	Yes	Yes	Yes	No
5 – Patterns, Relations, and Algebra	Yes	Yes	Yes	Yes
5 – Statistics and Probability	Yes	Yes	Yes	No
5 – Geometry and Measurement	Yes	Yes	No	No
Summary*	100%	100%	81%	13%

* This calculation is the total number of standards that met the criteria divided by the total possible number (4 standards x 4 levels = 16).

The MAPT for Math met the criteria for acceptable depth-of-knowledge consistency and balance of representation for all of the standards. The categorical concurrence requirements were met in 13 out of 16 standards (81%). The range of knowledge correspondence was the weakest dimension for the MAPT for Math. The requirement for this dimension was only met for one standard and was weakly met for two other standards. The low acceptable range of knowledge finding across the standards is due to the large number of objectives within each standard and the limited number of items. There is an average of 20 objectives per standard. To meet the criteria for the range of knowledge on average 10 objectives from each strand would need to each be assessed only once in a 40 item test. This is not realistic given that some standards are required to be more heavily weighted in the test specification table and that some objectives within a standard may require more than one item to fully assess that skill.

Reliability Results

The reliability of the ratings is judged through two analyses. Intraclass correlations were used to determine the reliability of the cognitive level classifications of the items. Webb et al. (2005) states that intraclass correlations should be .8 or greater and it is acceptable if it is .7 to .8. Average pairwise comparisons were used to determine the reliability of the item/objective matching. Webb (2005) suggested the average pairwise comparisons should be .6 or greater, but it is acceptable if it is .5 to .6. Table 11 lists the reliability results for each method for each assessment level.

Table 11 – Reliability Results

Test Level	Intraclass Correlation* (Item Cognitive Level Classifications)	Average Pairwise Comparisons** (Item/Objective Matching)
2	0.86	0.64
3	0.91	0.55
4	0.80	0.59
5	0.75	0.63

* Intraclass correlations: good > 0.8, acceptable between 0.7 and 0.8
** Average pairwise comparisons: good > 0.6, acceptable between 0.5 and 0.6

All reliability results met or exceeded Webb' s criteria for "acceptable" The acceptable levels of reliability will now be discussed in more detail. The intraclass correlation for level 5 was .75 which is an acceptable result based on Webb' s guidelines. This result is lower than the other levels most likely due to the increased complexity of the tasks at level 5. Mary discussed the increased complexity of the objectives at level 5 and stated, "I found [level 5] very difficult. Trying to figure out what objective it was going to was very difficult... It was difficult to determine what the items were measuring."

Two assessment levels had average pairwise comparisons in the acceptable range. The average pairwise comparison for level 3 was 0.55. Given the adaptive nature of the assessment, at level 3 participants were also likely to code items to objectives within level 2 and level 4 as well. The broader range of options was also true for level 4 and this also impacted the reliability of those matches (average pairwise comparison of 0.59). At level 4 participants were likely to code items to objectives from level 3 and level 5 as well. The lower level of agreement among participants' item-objectives matches for these two assessment levels is understandable. The detailed alignment results for each assessment level will be presented next.

Level 2 Alignment Dimensions

Level 2 met the alignment requirements for each dimension except categorical concurrence in Patterns, Relations, and Algebra and the range of knowledge correspondence for all four standards. Tables for each dimension and observations from these results are listed below.

Categorical concurrence. Table 12 lists the categorical concurrence results for level 2. Three out of four of the standards within level 2 met the criteria for acceptable categorical concurrence. This means that for those standards there is at least an average of six item/objective matches. The Number Sense strand has the highest average number of hits (greatest number of item/objective matches on average) while the Patterns, Relations, and Algebra strand has the lowest average number of hits. The distinction between these two strands can sometimes be confusing for participants. For example, some Number Sense objectives refer specifically to the operations (addition, subtraction, etc.), but then there is a Patterns, Relations, and Algebra objective referring to the mathematical signs for the operations (+, -, etc.). The distinction between when the item is asking about the skill and when it is asking about the symbol may not always have been clear to the participants. The Number Sense strand also has the largest standard deviation.

Table 12 – Level 2 Categorical Concurrence

Standards		Hits		Categorical Concurrence*
Title	Objs #	Mean	S.D.	
Level 2 Number Sense	19	14.17	3.24	YES
Level 2 Patterns, Relations, and Algebra	11	5	1.63	NO
Level 2 Statistics and Probability	15	9	1.29	YES
Level 2 Geometry and Measurement	15	11.17	1.57	YES

* "Yes" - mean number of hits is six or more.
"Weak" - mean number of hits is five to six.
"No" - mean number of hits is less than five.

Depth-of-knowledge consistency. Each standard within level 2 met the requirements for acceptable depth-of-knowledge consistency as shown in Table 13. Well over 50% of the hits within each standard at the test level are at a cognitive level at or above the objective to which the item is matched. This means that the items in the level 2 assessment are meeting or exceeding the cognitive level expectations as set forth in the objectives to which the items are matched.

Table 13 – Level 2 Depth-of-Knowledge Consistency

Standards		Hits		Level of Item w.r.t. Standard						DOK Consistency*
				% Under		% At		% Above		
Title	Objs #	M	S.D.	M	S.D.	M	S.D.	M	S.D.	
Level 2 Number Sense	19	14.17	3.24	12	27	56	40	32	41	YES
Level 2 Patterns, Relations, and Algebra	11	5	1.63	5	21	39	48	57	48	YES
Level 2 Statistics and Probability	15	9	1.29	19	39	38	44	44	45	YES
Level 2 Geometry and Measurement	15	11.17	1.57	17	35	52	45	31	43	YES

*"Yes" - 50% or more of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.
"Weak" - 40% to 50% of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.
"No" - less than 40% items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.

Range of knowledge correspondence. Level 2 did not acceptably meet the criteria for range of knowledge correspondence for any of the standards as shown in Table 14. This means less than 50% of the objectives within each level 2 standard were assessed. Forty-six percent of the objectives in standard Level 2 Number Sense had hits so this standard weakly met the criterion for range of knowledge correspondence.

Table 14 – Level 2 Range of Knowledge Correspondence

		Range of Objectives				Range of Knowledge*
		# Objs Hit		% of Total		
Standard	Objs #	Mean	S.D.	Mean	S.D.	
Level 2 Number Sense	19	8.83	1.07	46	6	Weak
Level 2 Patterns, Relations, and Algebra	11	3.67	1.25	33	11	No
Level 2 Statistics and Probability	15	4.5	0.96	30	6	No
Level 2 Geometry and Measurement	15	6	1	40	7	No

* "Yes" - 50% or more of the objectives had at least one item/objective match.
"Weak" - 40% to 50% of the objectives had at least one item/objective match.
"No" - 40% or less of the objectives had at least one item/objective match.

Balance of representation. Each standard within the level 2 assessment met the requirements for balance of representation as shown in Table 15. This means that of the objectives that are assessed, items are evenly dispersed among those objectives.

Table 15 – Level 2 Balance of Representation

		Balance Index		
Standard	Objs #	Mean	S.D.	Balance of Representation*
Level 2 Number Sense	19	0.8	0.05	Yes
Level 2 Patterns, Relations, and Algebra	11	0.92	0.08	Yes
Level 2 Statistics and Probability	15	0.72	0.08	Yes
Level 2 Geometry and Measurement	15	0.76	0.04	Yes

* "Yes" - Balance Index was .7 or above (items evenly distributed among objectives).
"Weak" - Balance Index was .6 to .7 (a high percentage of items coded to two or three objs).
"No" - Balance Index was .6 or less (a high percentage of items coded to one obj.)

Level 3 Alignment Dimensions

Level 3 also met the alignment requirements for each dimension except categorical concurrence in Patterns, Relations, and Algebra and range of knowledge correspondence for all four standards. Tables for each dimension and observations from these results are listed below.

Categorical concurrence. Three out of four strands met the requirements for acceptable categorical concurrence based on the Webb criteria as shown in Table 16. This means that the three standards meeting the criteria each had an average of at least six item/objective matches. Similarly to level 2, the Number Sense strand has the highest average number of hits while the Patterns, Relations, and Algebra strand has the lowest. Again, the distinction between these two strands can sometimes be confusing. The Number Sense strand again also has the largest standard deviation.

Table 16 – Level 3 Categorical Concurrence

Standards			Hits		Categorical Concurrence*
Title	Goals #	Objs #	Mean	S.D.	
Level 3 Number Sense	3	25	9.83	2.54	YES
Level 3 Patterns, Relations, and Algebra	4	14	5.67	0.94	NO
Level 3 Statistics and Probability	5	30	7.17	1.95	YES
Level 3 Geometry and Measurement	4	19	7	1.73	YES

* "Yes" - mean number of hits is six or more.
"Weak" - mean number of hits is five to six.
"No" - mean number of hits is less than five.

Depth-of-knowledge consistency. All of the standards within level 3 met the requirements for depth-of-knowledge consistency as shown in Table 17. Well over 50% of the hits for the standards within this level are at a cognitive level that is at or above the objective to which the item is matched.

Table 17 – Level 3 Depth-of-Knowledge Consistency

Standards			Hits		Level of Item w.r.t. Standard						DOK Consistency*
					% Under		% At		% Above		
Title	Goals #	Objs #	M	S.D.	M	S.D.	M	S.D.	M	S.D.	
Level 3 Number Sense	3	25	9.83	2.54	8	27	36	47	56	48	YES
Level 3 Patterns, Relations, and Algebra	4	14	5.67	0.94	12	31	81	38	7	25	YES
Level 3 Statistics and Probability	5	30	7.17	1.95	7	25	64	44	29	41	YES
Level 3 Geometry and Measurement	4	19	7	1.73	36	45	36	42	27	42	YES

*" Yes" - 50% or more of the items were rated as " at" or " above" the Depth-of-Knowledge level of the corresponding objectives.
" Weak" - 40% to 50% of the items were rated as " at" or " above" the Depth-of-Knowledge level of the corresponding objectives.
"No" - less than 40% items were rated as " at" or " above" the Depth-of-Knowledge level of the corresponding objectives.

Range of knowledge correspondence. None of the four standards within level 3 met the criteria for acceptable range of knowledge correspondence as shown in Table 18. This means that less than 50% of the objectives within each standard at level 3 are assessed.

Table 18 – Level 3 Range of Knowledge Correspondence

		Range of Objectives				Range of Knowledge*
		# Objs Hit		% of Total		
Standard	Objs #	Mean	S.D.	Mean	S.D.	
Level 3 Number Sense	25	6.50	1.61	26	6	No
Level 3 Patterns, Relations, and Algebra	14	4.83	1.07	35	8	No
Level 3 Statistics and Probability	30	5.00	1.29	17	4	No
Level 3 Geometry and Measurement	19	4.67	0.75	25	4	No

* "Yes" - 50% or more of the objectives had at least one item/objective match.
"Weak" - 40% to 50% of the objectives had at least one item/objective match.
"No" - 40% or less of the objectives had at least one item/objective match.

Balance of representation. Each standard within level 3 met the requirements for balance of representation as shown in Table 19. This means that, of the objectives that are assessed, items are evenly dispersed among those objectives.

Table 19 – Level 3 Balance of Representation

		Balance Index		
Standard	Objs #	Mean	S.D.	Balance of Representation*
Level 3 Number Sense	25	0.81	0.03	Yes
Level 3 Patterns, Relations, and Algebra	14	0.88	0.05	Yes
Level 3 Statistics and Probability	30	0.83	0.08	Yes
Level 3 Geometry and Measurement	19	0.86	0.08	Yes

* "Yes" - Balance Index was .7 or above (items evenly distributed among objectives).
"Weak" - Balance Index was .6 to .7 (a high percentage of items coded to two or three objs).
"No" - Balance Index was .6 or less (a high percentage of items coded to one obj.)

Level 4 Alignment Dimensions

Level 4 met the alignment requirements for each dimension except the range of knowledge for all four standards. Tables for each dimension and observations from these results are listed below.

Categorical concurrence. Each standard for level 4 met the requirements for categorical concurrence as shown in Table 20. This means the assessment adequately represents the content expressed in the ABE math standards given that each standard is represented by at least six item/objective matches.

Table 20 – Level 4 Categorical Concurrence

Standards			Hits		Categorical Concurrence*
Title	Goals #	Objs #	Mean	S.D.	
Level 4 Number Sense	3	29	12.83	2.19	Yes
Level 4 Patterns, Relations, and Algebra	4	17	11.5	2.57	Yes
Level 4 Statistics and Probability	5	30	7.67	1.70	Yes
Level 4 Geometry and Measurement	4	23	7.67	0.47	Yes

* "Yes"- mean number of hits is six or more.
"Weak"- mean number of hits is five to six.
"No"- mean number of hits is less than five.

Depth-of-knowledge consistency. Each standard for level 4 met the requirements for depth-of-knowledge consistency as shown in Table 21. Well over 50% of the hits within this level are at a cognitive level that is at or above the objectives to which the items are matched.

Table 21 – Level 4 Depth-of-Knowledge Consistency

Standards			Hits		Level of Item w.r.t. Standard						DOK Consistency*
					% Under		% At		% Above		
Title	Goals #	Objs #	M	S.D.	M	S.D.	M	S.D.	M	S.D.	
Level 4 Number Sense	3	29	12.83	2.19	12	33	47	49	41	48	Yes
Level 4 Patterns, Relations, and Algebra	4	17	11.5	2.57	32	44	41	44	27	41	Yes
Level 4 Statistics and Probability	5	30	7.67	1.70	7	25	40	46	53	47	Yes
Level 4 Geometry and Measurement	4	23	7.67	0.47	8	25	56	45	35	43	Yes

*"Yes" - 50% or more of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.
"Weak" - 40% to 50% of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.
"No" - less than 40% items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.

Range of knowledge correspondence. Level 4 did not meet the criteria for range of knowledge correspondence for three out of the four standards and only weakly met the criteria for the fourth standard as shown in Table 22. Forty-three percent of the objectives in standard Level 4 Patterns, Relations, and Algebra had hits so this standard weakly met the criterion for range of knowledge correspondence. Less than forty percent of the objectives in the other three standards are assessed.

Table 22 – Level 4 Range of Knowledge Correspondence

		Range of Objectives				Range of Knowledge*
		# Objs Hit		% of Total		
Standard	Objs #	Mean	S.D.	Mean	S.D.	
Level 4 Number Sense	29	9.33	1.49	32	5	No
Level 4 Patterns, Relations, and Algebra	17	7.33	2.21	43	13	Weak
Level 4 Statistics and Probability	30	4.83	1.07	16	4	No
Level 4 Geometry and Measurement	23	5.67	0.75	25	3	No

* "Yes" - 50% or more of the objectives had at least one item/objective match.
"Weak" - 40% to 50% of the objectives had at least one item/objective match.
"No" - 40% or less of the objectives had at least one item/objective match.

Balance of representation. Each standard within level 4 met the requirements for balance of representation as shown in Table 23. This means that, of the objectives that are assessed, items are evenly dispersed among those objectives.

Table 23 – Level 4 Balance of Representation

		Balance Index		
Standard	Objs #	Mean	S.D.	Balance of Representation*
Level 4 Number Sense	29	0.79	0.08	Yes
Level 4 Patterns, Relations, and Algebra	17	0.76	0.04	Yes
Level 4 Statistics and Probability	30	0.80	0.05	Yes
Level 4 Geometry and Measurement	23	0.83	0.03	Yes

* "Yes" - Balance Index was .7 or above (items evenly distributed among objectives).
"Weak" - Balance Index was .6 to .7 (a high percentage of items coded to two or three objs).
"No" - Balance Index was .6 or less (a high percentage of items coded to one obj.)

Level 5 Alignment Dimensions

Level 5 also met the alignment requirements for each dimension except categorical concurrence in Geometry and Measurement and the range of knowledge correspondence for three out of four standards. Tables for each dimension and observations from these results are listed below.

Categorical concurrence. Three out of four standards met the requirement for categorical concurrence as shown in Table 24. The standard for Geometry and Measurement did not meet the criteria as there was only an average of 4.67 hits for this standard.

Table 24 – Level 5 Categorical Concurrence

Standards			Hits		Categorical Concurrence*
Title	Goals #	Objs #	Mean	S.D.	
Level 5 Number Sense	3	17	6.67	1.49	Yes
Level 5 Patterns, Relations, and Algebra	4	11	9.67	1.49	Yes
Level 5 Statistics and Probability	5	24	11.33	2.75	Yes
Level 5 Geometry and Measurement	4	14	4.67	1.11	No

* "Yes" - mean number of hits is six or more.
"Weak" - mean number of hits is five to six.
"No" - mean number of hits is less than five.

Depth-of-knowledge consistency. Each standard for level 5 met the requirements for depth-of-knowledge consistency as shown in Table 25. Well over 50% of the hits for the standards within this level are at a cognitive level that is at or above the objective to which the item is matched.

Table 25 – Level 5 Depth-of-Knowledge Consistency

Standards			Hits		Level of Item w.r.t. Standard						DOK Consistency*
					% Under		% At		% Above		
Title	Goals #	Objs #	M	S.D.	M	S.D.	M	S.D.	M	S.D.	
Level 5 Number Sense	3	17	6.67	1.49	12	33	54	48	34	45	Yes
Level 5 Patterns, Relations, and Algebra	4	11	9.67	1.49	26	40	46	46	28	42	Yes
Level 5 Statistics and Probability	5	24	11.33	2.75	18	37	47	47	35	46	Yes
Level 5 Geometry and Measurement	4	14	4.67	1.11	27	43	42	47	31	45	Yes

*"Yes" - 50% or more of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.
"Weak" - 40% to 50% of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.
"No" - less than 40% items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.

Range of knowledge correspondence. Level 5 did not meet the criteria for range of knowledge for three out of four standards as shown in Table 26. The only standard to assess over 50% of the objectives within a standard was the Patterns, Relations, and Algebra standard. This standard only had 11 objectives so the 50% goal was more easily attainable.

Table 26 – Level 5 Range of Knowledge Correspondence

		Range of Objectives				Range of Knowledge*
		# Objs Hit		% of Total		
Standard	Objs #	Mean	S.D.	Mean	S.D.	
Level 5 Number Sense	17	5.50	1.26	32	7	No
Level 5 Patterns, Relations, and Algebra	11	5.83	0.90	53	8	Yes
Level 5 Statistics and Probability	24	7.17	1.07	30	4	No
Level 5 Geometry and Measurement	14	4.00	1.15	29	8	No

* "Yes" - 50% or more of the objectives had at least one item/objective match.
"Weak" - 40% to 50% of the objectives had at least one item/objective match.
"No" - 40% or less of the objectives had at least one item/objective match.

Balance of representation. Each standard within level 5 met the requirements for balance of representation as shown in Table 27. This means that, of the objectives that are assessed, items are evenly dispersed among those objectives.

Table 27 – Level 5 Balance of Representation

		Balance Index		
Standard	Objs #	Mean	S.D.	Balance of Representation
Level 5 Number Sense	17	0.88	0.10	Yes
Level 5 Patterns, Relations, and Algebra	11	0.80	0.06	Yes
Level 5 Statistics and Probability	24	0.78	0.02	Yes
Level 5 Geometry and Measurement	14	0.90	0.07	Yes

* "Yes" - Balance Index was .7 or above (items evenly distributed among objectives).
"Weak" - Balance Index was .6 to .7 (a high percentage of items coded to two or three objs).
"No" - Balance Index was .6 or less (a high percentage of items coded to one obj.)

Source of Challenge Comments

As the participants coded each item to a cognitive level and up to three objectives, they also considered if the item had a source of challenge issue. There were 36 source of challenge comments across the four levels of the MAPT for Math. The specific source of challenge comments and how they were coded appears in Appendix F. The assessment for each level was 40 items and the percentage of items noted with a source of challenge in each level was: 3% in level 2, 18% in level 3, 10% in level 4, and 8% in level 5. Of the 36 total source of challenge comments, 3% (n=1) caused one item to be changed prior to the test becoming operational, 44% (n=16) did not require any items to be revised, and 53% (n=19) pointed out issues with items that should be examined for possible changes in future versions of the MAPT.

Of the 16 comments that do not require future modifications, six of comments were related to the level where the item appeared. The participants noted if they thought the item were too easy or difficult for that learning level. Based on the item statistics from the pilot data, the placement of the items are correct and do not need to be adjusted. This may be revisited when additional operational data are available, but for now it does not need to be examined. Six of the comments also noted concerns about the specific skill required but these are skills that are specified in the curriculum frameworks so they are relevant. Three of the comments were more general comments about the item rather than specific source of challenge issues. And one comment suggested a change that would add unnecessary information to an item.

The 19 items that need to be more closely examined to determine if modifications are required in future versions of the MAPT were coded to determine the

types of changes that might be required. Of these 19 source of challenge participant comments, eight addressed concerns that the vocabulary used might be unfamiliar to students who are English language learners. Five comments noted concerns that the context of an item might be unfamiliar for students who are English language learners. Three comments had concerns about a graphic being unclear. Two comments expressed concern about the placement of the calculator. The calculator currently appears between the question and the answer choice for short text items. These comments noted that this placement might be difficult for students how have reading issues. The final comment noted that the label for a thermometer should be spelled out instead of abbreviated to be consistent with the answer choices.

All items were reviewed by a Sensitivity and Bias committee to ensure the items did not unduly favor or harm any group of students. The participants in this study, however, still found issues of concern regarding English language learners and students with reading disabilities. The Source of Challenge notes are a helpful way to revisit these concerns and make changes for future versions of the MAPT.

General Participant Comments

While matching items to objectives, the participants noted any general comments they had about an item. There were 180 general comments made by the participants across all of the levels (Appendix G). The comments were coded and divided among six themes.

The most prevalent theme of comment related to specific framework observations or recommendations. There were 64 comments within this category (36%). These comments noted how some objectives were too specific or not specific

enough, and also that some skills could be covered earlier. For example, line graphs do not get introduced until level 4, or miles was not included in the measurement objective. Comments were also made about possibly missing objectives such an objective specifically about horizontal line meaning division or exponential growth in patterns. Finally, framework comments touched on the need to reorder some of the objectives. For example, one participant thought that objective 4P-3.9 should be at level 3 and the objective about prime numbers should come earlier than level 5. This theme of frameworks observations among the general comments illustrates how the alignment review process can inform future framework revisions as teachers work more closely with the objectives and think about how they are operationalized in assessment items and relative to their classroom instruction.

The second most prevalent theme among the general comments was comments that showed the participants were unsure of the item-objective match. There were 56 comments that were coded in this category (31%). There were four categories of comments within the theme of comments related to item-objective matches. First, forty-six percent (26) of the unsure of match comments were due to concerns about how well the item fit with the associated objective. For these comments the participants were concerned that the item only measured a part of the objective to which it was matched. Partial item-objective matching was highlighted as a concern that is captured through the Achieve approach but is not a distinction made in the Webb model (Rothman, 2003). Other comments noted concerns that the item asked more of the student than what is literally stated in the objective, but this was the best fit possible. These types of comments that focused on questions regarding the item-objective match

highlight the concern that the Webb method does not allow participants to discuss the quality of the match between the item and the objective.

Second, thirty percent (17) of the "unsure of match" comments had to do with difficulty finding an objective to match the skills required in the item. The table format helped to facilitate the review process, but a number of participants commented in the focus group that using the computer for the item-objective process was much easier. Given the number of objectives, searching through the tables looking for specific objectives was sometimes difficult.

Third, fourteen percent (8) of the "unsure ofmatch" comments had to do with just general uncertainty by the participants. These eight comments stated in general that the participants were "not really sure" or they "didn' t know."It was unclear from these comments if the trouble was in understanding what the item was asking or in finding an objective to match the skills of the item. These types of comments illustrate that a larger discussion component in the item-objective matching process could have been helpful.

Fourth, the remaining five comments (9%) within "unsure of match"were participants' notes about why they selected the item-objective match they made. These types of comments also showed that participants could have benefited from more discussion time to share their rationale behind their matches and learn from each other.

The third theme among the general comments was 27 comments (15%) that had item critical comments. These comments offered specific feedback about how to improve an item or expressed concerns at the level at which the item was appearing. The comments will be helpful in revisiting the items for future versions of the MAPT to

suggest possible modifications. Again, as noted in the source of challenge comments, in terms of level placement, the placement is determined based on the pilot data for that item.

The fourth theme among the general comments was 19 comments (11%) where the participants expressed uncertainty about the cognitive level of the item. For example, at level 5, participants noted seven items that were difficult to distinguish between App and ASE. This confusion highlighted the potential value of increased discussion in the item-objective matching and coding process as participants could share their ideas and learn from each other.

The fifth theme among the general comments was 13 comments (7%) that were just general feedback about an item and did not require additional attention. And the sixth theme was one comment (1%) about the functionality of the calculator.

The general comments illustrate the connections the alignment process can have to future modifications for both the MAPT for Math and the Math ABE standards. They also show the importance of providing a way for participants to express their thoughts during the coding process. In the absence of any in-depth opportunities for discussion, the participants used the comment field as a way to talk about their thinking through the item-objective matching and coding process.

Debriefing Questions

The debriefing survey question results were reviewed holistically to understand general themes. Then specific results for each assessment level are summarized. There were differences about the degree and quality of the alignment across all of the levels between the quantitative Webb results and the qualitative comments shared by the

participants through the debriefing questions. While each assessment level of the MAPT for Math met the requirements for balance of representation, the participants' responses to the debriefing questions all highlighted topics within each level that were over or underrepresented. While the Webb analysis looks at the data from the perspective of standards and objectives, the participants responded by noting specific topics (such as shapes, number lines, inequalities) that they thought were missing. Each assessment level also met the requirements for acceptable depth-of-knowledge consistency. At each level, however, the participants noted cognitive areas they thought were lacking causing the assessment to not be as challenging as they thought it should be.

Level 2. While level 2 met the requirements for acceptable balance of representation, the participants listed specific topics they thought were over or under represented. Participants thought there were too many items related to time, patterns, and graphs/tables. Instead participants wanted to see more basic operations, calculator usage, symmetry, shapes, number lines, inequalities, and missing variables. While the range of knowledge correspondence results for this level did show that there were not a large enough percentage of the objectives that were assessed, it did not help to specifically show the types of items that are missing. Participants' debriefing answers assisted in identifying these underrepresented topics. Although level 2 met the requirements for depth-of-knowledge consistency, the participants wanted to see more Knowledge and Comprehension and Analysis, Synthesis and Evaluation items at this level.

Level 3. This level also met the requirements for acceptable balance of representation but the participants still listed topics they thought were over or under represented. Participants thought there were too many items about graphing and pulling information from graphs. Then they listed a number of specific topics they would like to see in the future. These topics included: specific statistics and probability items, algebra, squares/cubes, rounding, solving expressions, mean/median, symmetry, triangles/angles, measurement, and order of operations. Again the range of knowledge correspondence dimension showed that there were not a large enough percentage of the objectives assessed, however, it did not help to specifically show the types of items that are missing. The participants' debriefing comments were helpful to show the types of items they would like to see in future versions of the assessment. Although level 3 met the requirements for depth-of-knowledge consistency, the participants also noted they would like to see more Analysis, Synthesis, and Evaluation items instead of Application items. Participants did think there was a good blend of Knowledge and Comprehension items.

Level 4. This level also met the requirements for acceptable balance of representation but the participants still listed topics they thought were over or under represented. The participants thought there were too many items with charts. Then there were a number of topics the participants would like to see assessed more. These topics included: finding percentages, circles, fractions, number lines, and inequalities. Again, although the range of knowledge correspondence dimension showed that there were not a large enough percentage of the objectives assessed, it did not help to specifically show the types of items that are missing. The participants' debriefing

comments were helpful to show the types of items they would like to see in future versions of the assessment. Although this level met the requirements for depth-of-knowledge consistency, the participants noted there were more Analysis, Synthesis, and Evaluation items at this level than at the earlier levels but there still seemed to be too many Application items. One participant also requested more Knowledge and Comprehension items at level 4.

Level 5. This level also met the requirements for acceptable balance of representation but the participants listed topics they thought were over or under represented. The participants thought there were too many Statistics and Algebra items. Unfortunately this type of comment did not help to inform the specific types of items within each of these broad strands that might have been overrepresented. There were a number of topics the participants would like to see assessed by more items. These topics included: Number Sense (again a broad strand), Geometry (angles, triangles), area/perimeter/volume, symmetry, fraction/proportion. This level did not meet the requirement of categorical concurrence for the standard Geometry and Measurement so it is helpful to see that the participants thought there should be more angles, triangles, symmetry, and items related area/perimeter/volume. Level 5 also met the requirements for depth-of-knowledge consistency and the participants thought there was a better balance of the cognitive levels at this level. Participants would still like to see more Analysis, Synthesis, and Evaluation items and less Application items.

Summary evaluation. At the end of the debriefing questions the participants were asked: " What is your general opinion of the alignment between the standards and the assessment?" The results for each level are presented in Table 28. One participant

was very concerned with the level 2 assessment. In the debriefing questions this participant wrote that there were not enough Knowledge and Comprehension items at this level to assess students' basic computation understanding. Beyond that one participant, the results for this survey question show the participants thought the assessments were acceptably aligned or required slight improvements.

Table 28 – Participants' Summary Evaluation Regarding the Degree of Alignment

	Assessment Levels			
	2	3	4	5
Perfect alignment	0%	0%	0%	0%
Acceptable alignment	67%	50%	67%	67%
Needs slight improvement	17%	50%	33%	33%
Needs major improvement	17%	0%	0%	0%
Not aligned in any way	0%	0%	0%	0%

One major concern that participants had for each level was the topics that were not assessed in the assessment levels the participants saw. The debriefing questions highlight specific topics that the participants would like to see in the future. From the teachers' commentsit was apparent that they thought about assessment at a topical level (number lines, estimation, shapes, etc.) rather focusing on the specific objectives. The topical view of the data is masked by the more specific objective view or the more general standard view used in the Webb methodology criteria calculations. Additionally, participants noted they would like to see more Analysis, Synthesis, and Evaluation items at each level. Table 29 shows the number of items that were reclassified from the level they were originally written for to a different cognitive level in the alignment process. Only items where 4 or more participants agreed with the reclassification are shown. The largest number of items are reclassification from

Analysis, Synthesis, and Evaluation to Application (15 items). The distinction between cognitive levels can be difficult to determine and will need to be evaluated more in the future to ensure the MAPT for Math has items that adequately represent that cognitive level.

Table 29 – Revised Item Classifications

	Cognitive Level from Alignment Process			
Original Cognitive Level of the Item	KC	App	ASE	Total
KC		3		3
App	7			7
ASE	2	15		17
Total	9	18		27

Alignment to the Test Specifications

A key component of the test development process is the determination of the test specifications. This is the document that connects the standards and the assessment as it sets out the relative emphasis for each strand and cognitive level. The test specification table for the MAPT for Math was presented in Table 2.

In the Webb methodology, the results of each participant' s coding was averaged to calculate the results for each alignment dimension. As noted earlier then, if two participants coded an items as measuring Number Sense, and two participants coded the same item as measuring Patterns, Relations, and Algebra, and the last two participants coded the same item as measuring Geometry and Measurement, all results are included and averaged. The implication is that an understanding of what the item is truly measuring, even at the strand level, is masked. To see the impact of setting a minimum level of participant agreement, the criteria used in traditional content validity studies was applied to the data gathered using the Webb methodology. In this analysis only

items that had 4 out of 6 or more participants agreeing on the strand (Number Sense, Patterns, Relations, and Algebra, etc.) or cognitive level were included. Then the results of this analysis were compared to the original test specifications to determine how well the test is actually measuring what it was designed to measure. This gave a more accurate view of whether the test is accomplishing what it was designed to do than the Webb methodology, which sets the same criteria for each standard at each level (6 items per standard and 50% of the objectives measured by items that are at or above the cognitive level of that objective).

Table 30 shows the percentage of items classified by a minimum of 4 out of 6 participants for each strand and cognitive level for each level of the test (Actual). This is then compared to the original test specification proportions (Target). The difference is also calculated. The results show that no strand has more than plus or minus five percentage points relative to the target goals. For a 40 item test this translates to 1 or 2 items that need to be adjusted among the strands at each of the levels. This finding supports the categorical concurrence results using the Webb methodology where the majority of the standards met the requirements for acceptable categorical concurrence. However, the cognitive areas show greater discrepancies. One particular area of concern is at level 2 where the goal was to have 41% of the test represent Knowledge and Comprehension but there was agreement that only 28% came from this level. We should have at least five more items written to the Knowledge and Comprehension level. This was supported through the debriefing comments where participants stated they wanted to see more straight computation problems at the lower level as this is a

skill students struggle with. At each test level the percentage of Analysis, Synthesis, and Evaluation items needed to be increased.

Table 30 – Participant Agreement Criteria Compared to Test Specifications

		Cognitive Level						
Level	Strand	KC	App	ASE	Split	Actual	Target	Difference
2	N	15%	18%	0%	0%	33%	35%	-3%
	P	3%	5%	3%	0%	10%	15%	-5%
	S	3%	8%	0%	13%	23%	25%	-3%
	G	8%	15%	0%	5%	28%	25%	3%
	Split	0%	3%	0%	5%	8%		
Actual		28%	48%	3%	23%	100%		
Target		41%	41%	18%				
Difference		-14%	7%	-16%				
3	N	10%	23%	0%	0%	33%	30%	3%
	P	8%	8%	0%	0%	15%	20%	-5%
	S	3%	10%	8%	3%	23%	25%	-3%
	G	13%	10%	0%	3%	25%	25%	0%
	Split	3%	3%	0%	0%	5%		
Actual		35%	53%	8%	5%	100%		
Target		35%	45%	20%				
Difference		0%	8%	-13%				
4	N	13%	8%	0%	5%	25%	25%	0%
	P	5%	15%	0%	5%	25%	25%	0%
	S	3%	10%	3%	5%	20%	25%	-5%
	G	8%	13%	0%	3%	23%	25%	-3%
	Split	0%	8%	0%	0%	8%		
Actual		28%	53%	3%	18%	100%		
Target		25%	50%	25%				
Difference		3%	3%	-23%				
5	N	8%	10%	0%	0%	18%	15%	3%
	P	3%	13%	3%	8%	25%	30%	-5%
	S	5%	13%	5%	13%	35%	30%	5%
	G	3%	10%	0%	8%	20%	25%	-5%
	Split	0%	0%	3%	0%	3%		
Actual		18%	45%	10%	28%	100%		
Target		18%	41%	41%				
Difference		-1%	4%	-31%				

The fact that the test specifications were not incorporated into the alignment methodology also influenced the degree to which distinctions between the standards could be made. For example, while the number of hits at level 2 might seem high (14.17) compared to Patterns (5) these different emphases may be what is required in

the test specifications document. The focus group discussion looked at the results of the Webb alignment methodology analysis compared to the requirements set forth in the test specification table.

Beth noted in the focus group discussion that she was pleased to see Number Sense had the highest number of hits. She stated, "That doesn' t concern me. Level 2 should be heavy in Number Sense. Thirty-five percent should be Number Sense. Thirteen items." It was important for Beth to compare the Number Sense to the goals set out in the original test specifications document. As Beth notes, the majority of the items for level 2 should come from Number Sense. Melissa would like to see even more Number Sense items and thought the percentages in the test specification table should be adjusted.

Beth also emphasized the importance of revisiting the test specifications document after completing this alignment process. Now that the participants saw how a test was operationalized with forty items, they thought it might be important to revisit the percentage distributions set forth in the test specification table. Beth stated, "On the percent distribution, this was developed day 1 [back in] 2003. We sat there [from] 8-3 and at some point in the afternoon [we discussed] what percent should be distributed at each level. And this hasn' t been revisited. Do you think it is a good distribution at the levels? Is there anything that you think should be changed?" Melissa replied, "I might change the Statistics down. I thought there were so many." And this comment was supported in the debriefing comments analyzed above where Statistics was often seen as overrepresented at each level.

Len supported Melissa' s point and stated, " Some of the newer stuff at level 5 are advanced geometry and algebra. Maybe switch Statistics and Geometry at level 5." Through the process of comparing the categorical concurrence and depth-of-knowledge results from the Webb methodology to the test specification table and the group discussion, the participants saw that different strands might need more emphasis than others and the test specification table is the place to make these requirements known for test development. Beth concluded by saying, " Perhaps before year 2 this should be revisited." This is an important point as the test specification is the foundation for how the test gets constructed and how pilot testing evolves. However, the Webb methodology, with its criteria of six items per strand and 50% or more of the objectives assessed by items that are at or above the cognitive level of the objective does not allow for differentiation among the standards or cognitive levels.

Results for the Second Research Question - How does teachers' involvement in the alignment process influence their views of the standards, the assessment, and their approach to instruction?

The second research question explored how teachers' involvement in the alignment process influenced their view of the standards, the assessment, and their instruction. The data from the ongoing group discussions and the focus group discussion were analyzed to answer this research question. The themes related to the importance of discussion will be discussed first because the findings span the three aspects of the alignment question. Then the themes for each of the alignment components, standards, assessment, and instruction, will be presented.

Importance of Discussion

Throughout this alignment study the participants appreciated opportunities to talk with their colleagues about what they were doing and thinking as a means to more fully understand and be a part of the alignment process. After participating in each alignment activity the participants shared their thoughts with their colleagues, discussed concerns they had, and shared their passion for the topic as they sought to understand or improve the components. During the objective coding process (phase 1), the participants were required to interact to develop a consensus view of the cognitive level for each objective. While the item-objective matching and coding process (phase 2) was a more independent activity, the participants enjoyed the training, informal opportunities to share ideas, and our closing discussions. Finally, during the focus group the participants had the opportunity to share specific thoughts about how the alignment process evolved, what they would change, and what they learned from the

process. There are three themes within the importance of the discussions to the participants. First, the discussions facilitated the process of reaching consensus about how the objectives should be viewed. Second, the discussions ensured a common understanding of terms and ideas. Third, the discussions helped the participants to feel validated about their understanding of the standards and the assessment.

Facilitated the process of determining consensus. The first step of this alignment study required the participants to determine the cognitive level required to accomplish each of the objectives in the Math ABE standards. This was a discussion rich process where each objective was rated and then the objectives that were not unanimously classified were discussed. Through the discussions the participants mined the language of the objective, posed potential items that could address the objective, and compared the requirements of the objective to other objectives where consensus had been achieved. Throughout the discussions the participants learned more about the objectives and the cognitive level distinctions as they shared their ideas and experiences with each other.

Of the 313 objectives, 64% of the objectives required discussion to reach a consensus. For these 199 objectives, 67% (134) of the objectives had a consensus rating that agreed with the original majority viewpoint. For 27% (54) of the objectives the participants were originally split on how to rate the objective. And for 6% (11) of the objectives the final consensus rating was originally the minority viewpoint. Just going with the majority viewpoint, and not having a discussion, could have resulted in incorrect ratings for 65 objectives. The majority of the discussion focused on distinguishing between Knowledge and Comprehension and Application. Whereas

participants felt they had a good understanding of when objectives were requiring Analysis, Synthesis, and Evaluation skills, the distinction between Knowledge and Comprehension and Application was not always as clear. Often the participants saw verbs that looked like basic recall skills but then could see it required in a context. Of the 54 objectives that were split, 27 were evenly split 3-3. The majority of this discussion (21 out of 27 objectives) focused on whether objectives were Knowledge and Comprehension or Application.

After the task, the participants debriefed about the process and then we had follow up emails about the process. The participants all emphasized how much they liked the discussion and learning from each other. The participants realized they brought different backgrounds and experiences to this task and appreciated learning from each other. Although one participant expressed frustration at times with another participant, she valued the discussion process and learning from others. Another participant claimed to not be a "math expert" but she thought it was very important to discuss the way the objectives were expressed and how they were operationalized in items from a literacy perspective. Learning to listen to each other and see other perspectives helped all of the teachers to grow through the experience.

Sabrina stated that the discussion about the cognitive levels was "the most interesting part of the objective coding process." Beth further stated, "I also very much enjoyed the opportunity to reflect on the questions with colleagues whose opinion I respect." And Melissa talked about what she learned from others through the objective coding process. She stated, "I liked the part that we had to have consensus. When I was judging the frameworks, I was trying to think of a question that could be made up to test

that item. I often saw things a little differently from the others." Melissa even wanted to talk more about the process after it was concluded so a longer debriefing period might have been helpful. This stage of the alignment process was the most discussion intensive and, while time consuming, the participants enjoyed learning from each other and sharing ideas.

Common understanding of terms and skills. Participants worked together to ensure they had a common understanding of terms within the cognitive levels and the skills required by the objectives. This second theme illustrated how discussions were important to understand and debate differing opinions. For example, Len had an issue with the way the cognitive levels were grouped. He thought that combining Comprehension with Knowledge did not adequately represent the comprehension skills and this is an area his students struggle with in his classroom. During the focus group he stated,

> "I just want to make sure it gets on the videotape that I do think it is a mistake to bundle Knowledge and Comprehension. I think you are giving short shrift to comprehension and this is a big issue for my students. They actually write that on their evaluations at the end of the week ' I need to work on my reading comprehension.' And there are a lot of pieces to comprehension. And I think when we were doing the rating we were taking a lot of things that should have been comprehension and lumping them in with knowledge and not ever putting them ever in application. Ideally we should have all of the levels represented. If we need three then have Knowledge, then Comprehension/Application."

By combining Knowledge with Comprehension the group determined through discussions in the objective coding process that this did, in effect, say that this cognitive level was anything that tested a rote understanding or was out of a context. The participants thought that then when a skill was tested in a context it became Application. It could be helpful in the future to have the groupings as Knowledge,

Comprehension/Application, and Analysis/Synthesis/Evaluation. But allowing for this discussion during the objective coding process helped the debate over the cognitive level groupings to come into fruition.

Through discussions during the objective coding process, the group helped Len to see the distinction in the way the levels were currently defined. The participant was still adamant that the separation of levels was not what he would like to see and I think this influenced his rating of items during the item-objective matching process. Without the discussion during this phase he might have been more apt to rate items others saw as Application as Knowledge and Comprehension given his strong thoughts about the way those cognitive levels should be grouped.

Participants thought that increased discussion during Phase 2 could have helped in determining what the items were really measuring. Judy noted that she had difficulty matching some items, particularly at the higher levels, to objectives and was concerned that this would be a negative interpretation of the quality of the item. She stated, " Which is kind of too bad [that items without agreement might not be included]. Some of the items I looked at I thought this is a really good question. I don' t know what it is exactly but it is a really good question." Mary also supported the difficulty of determining what the items at the higher levels were measuring. She stated, " I found [level5] very difficult. Trying to figure out what objective it was going to was very difficult... It was difficult to determine what the items were measuring..In level 2 you could look at a question and say that is clearly measuring that objective. The language of the objectives could have been related." At level 5 there was a higher level of disagreement about what the items were measuring. If the item-objective matching

process included more discussion it could help the teachers to learn more about the objectives and how they were operationalized.

Deeper understanding and validation from others. Finally, through the discussion process during both the objective coding and the item-objective matching, the participants shared and learned from each other about their views of the standards and the assessment. The participants specifically commented on the value of these discussions and noted that this was a missing component in some of the activities. Judy stated, "I found all of the discussions where we had to agree or not agree very helpful. It made me look at the levels or items in a different way, and think to see what it really is testing." Although the item-objective matching process did not have an in-depth discussion component, the participants expressed appreciation for the times they could share ideas and wished they had more time to collaborate. In the training process of the item-objective matching, the participants shared ideas about the types of items that could match different objectives, how an item could be extended to match a different cognitive level, and how they worked with the table layout of the objectives. They seemed to really appreciate hearing each other' s ideas even if it did not make them change their original match.

During the focus group the participants shared how valuable they found the discussion process that was integral to the objective coding process. Judy stated, "I thought [the discussion] was very valuable because I wasn' t always all that positive that what I wanted was what I put down. So it was really helpful to hear what everyone else was thinking and then make a final decision on that. I thought it was very valuable." Sabrina also found the discussion process very helpful. She highlighted how it forced

her to interact with and really think about the frameworks. She stated, "I thought the objective coding process was great. I had read them but reading them and using them are two different things. I thought that was worthwhile."

Whereas in the objective coding process, the participants could learn from each other and feel validated about their ratings, in the item-objective matching process the participants felt very isolated. During both the item-objective matching process and the focus group discussion, a number of participants expressed concern that their ratings were way off and "out there," that they didn' t want to be an example. They were concerned that they were "wrong" even though it was discussed that there was no right answer. Judy stated, "It was also challenging to do the parts on my own. I felt like I was getting it all wrong." Melissa agreed with this feeling. Sabrina added that she "liked the part where we checked in, talked about consensus." Judy noted that when they were able to check in with each other she learned from the other participants about different ways to look at items. She stated, "As we were going through I just wished I could talk to someone about it because I just wasn' t sure about this one. Or when people would present a different thought I would think that' s right I didn' t think of that before." The participants would have liked more discussion in the item-objective matching process to build on the validation and support they found through the objective coding and focus group discussion pieces of the alignment process.

Math ABE Standards

The participants' view of the standards was influenced by the work they did in determining the cognitive level of each objective and the process of matching items to objectives. Through the objective coding process, the participants learned about the

cognitive expectations of the frameworks and gained a greater familiarity with the frameworks. This understanding of the frameworks was then augmented as they could see how those objectives are operationalized through the items to which they are matched. There were two themes within this area of the results. First, the table layout of the Math ABE standards was very helpful to the participants. This new layout grouped objectives by common topics and illustrated how the skills developed across the learning levels. The new table layout of the standards fostered a deeper understanding of the Math ABE standards among the participants. Second, the participants shared many ideas to modify and improve the Math ABE standards. Building on their work in the objective coding process, and the item-objective matching using the new layout of the Math ABE standards, the participants were then able to suggest changes to the frameworks that built on their deeper understanding of the Math ABE standards.

Table view of the Math ABE standards. The table view of the standards, with objectives grouped across the levels by common topics, helped the participants to find objectives, understand how this skill developed across the learning levels, and identify what skills might be missing or need to change in the Math ABE standards. The Math ABE standards released by ACLS is a 100+ page document listing different objectives grouped under different standards within different strands. This format was overwhelming to participants as they waded through the list to see what was included. When the participants began the item-objective matching process they expressed the value in the table layout of the Math ABE standards, a more manageable 20 page document. The table format summarized many details into a meaningful presentation.

Judy stated, "I think we teach by topic, not by individual objectives. I think you want to cover main topics as well as objectives. I think that will be more valuable."

The table format also helped to illustrate gaps in the frameworks. Len noted, "In the table view you can see some things that aren' t assessed until level 5 or are assessed at level 2 and at level 4 but at level 3 there is this hole." For example, prime numbers are not introduced until level 5 and Multiplication is addressed at level 2 and level 4 but does not have a specific objective at level 3. When the objectives were just a list within the Math ABE standards, these types of observations and conclusions were not possible. The table format, however, should not replace the list format of the curriculum frameworks. The latter format provides more information about the enabling skills required for each objective. These are skills that are components of the broader objective. One participant often used this as a reference point when talking about the item-objective matches.

Participants also thought the computer should have been available for all of the item-objective matching process to facilitate the search for item-objective matches. The participants did levels two and five independently as homework where they did have access to the electronic version of the table view of the standards. The group discussed how the computer could be used to search for key words in the table view of the objectives. Mary noted how helpful this was and stated, "For next time :-)...I found it easier to use the 'find' on the computer to look up standards and I think I was able to be more thorough searching for the correct fit." Beth also stated, "I did Level 5 with the print-outs, but did Level 2 electronically. I had never thought about that until you mentioned it yesterday. It is soooo much easier. I would say do it in a computer lab

and demonstrate how to do it this way. People who haven't used the feature before would resist it---but once they try it, I expect they will like it." While there were improvements to be made in terms of the table layout of the frameworks, the participants all agreed it facilitated their understanding of the objectives across the different levels and how they were matched to different items.

Math ABE standard modifications. This alignment study helped the participants to better understand the Math ABE standards (also referred to by the participants as "the frameworks), how they were created, and how they might be improved. Having a participant of the Math ABE standards development team, Beth, involved in the alignment process helped everyone to better understand differences between the different strands. Len noted, "And the frameworks, [the alignment process] really humanized them to me. The frameworks no longer appeared like something from on high from the DOE that you must obey. It was a document that went through a process, evolved, changed, and different people had input. It is what it is. I see it as more organic." Beth noted that the Statistics and Probability strand was much more specific and repetitive than the other strands. She stated, "I' m even thinking I worked on the statistics with someone who was loading the objectives and I think it could have been condensed by topic much better." Measurement terms also need to be examined to ensure they are consistent and build across the levels. Specific changes to the Math ABE standards noted by the participants are listed in Appendix H. What follows is a summary of the types of revisions the participants noted.

During the item-matching process the participants made 65 general item specific comments that related to the standards. These comments noted how some objectives

were too specific and there were some skills that should be covered earlier. For example, line graphs do not get introduced until level 4, multi-step problems were not introduced until the higher levels, and prime numbers are not introduced until level 5. These comments were also supported in the focus group discussion. Regarding line graphs, Judy agreed they should be introduced earlier. She stated, " We use line graphs to show a lot of information. You may be able to not read it completely but still be able to do something with it." And Melissa was troubled by the late introduction of multi-step operations. She stated, " There was nothing that said you can do multi-step problems in the frameworks [at level 2]. I was at a loss as to where to put that." These types of comments, both from the item-objective matching process and the focus group discussion, illustrated the importance of revisiting the frameworks now that it is apparent how they are operationalized in items and teachers have had more time to work with them thinking about how it relates to their classroom instruction.

Participants also noted gaps in the frameworks. At level 2 Melissa stated, " I can't believe the frameworks don't go into more specific tasks like addition of basic sums to 10, to 20, two digit with carrying, subtraction with and without borrowing, subtraction when zeros are involved, etc." At level 3, the participants noted there were no objectives associated with scale/proportion skills and at level 4, the participants would like to see scatterplots and pictographs included as assessed objectives. The participants also noted some objectives may be too specific. For example, a number required a calculation to be done in a real life context or that percentages use " friendly numbers." Finally, comments were made about possibly missing objectives such an objective specifically about horizontal line meaning division or exponential growth in

patterns. These comments illustrate how the participants' view of the Math ABE standards was influenced by their participation in this alignment study.

MAPT for Math

Through the alignment process the participants developed a greater understanding of the MAPT for Math. There were three themes within this area of the results. First, participants learned about the test construction process in general as the different steps of the alignment process helped the participants to see how the assessment was developed. Second, the participants used the topical terms to think about what was over or under assessed and they appreciated see alignment data using that framework. Third, while the alignment process was meant to help the participants see how the assessment aligned with the standards, they did have trouble stepping back to view the assessment as a whole after they concluded the item-objective matching process. Each of these themes will now be discussed.

Appreciation of test construction process. Given that the MAPT for Math is a new assessment, participation in this alignment study allowed the participants to gain an appreciation for how the test was developed from the beginning to this point. The participants enjoyed hearing about how the frameworks were created, how the items were written to frameworks, and how their work here was an important part of the test development process. During the focus group discussion, Beth asked about how items were selected to become operational. We discussed how it was based on performance on the pilot tests but also on the match to the content specifications developed for the test. Beth was part of the original committee that developed the test specifications.

After participating in this alignment process she thought it would be important to revisit the test specifications to ensure they are the correct proportions.

Mary also commented on how helpful it was to learn more about how the test got constructed in general. She stated, " I thought the whole process of looking at the items, rating the items, reading the items, checking the answers. Understanding everything that goes into it. I just thought someone sat down and said here' s 40 questions. Really I had no idea what the process was for something like this." The teachers began to see how the assessment connected to the frameworks and can connect to what occurs in their classroom.

Topical view of the data. The results of the participants' debriefing questions and their discussions prior to the focus group showed me that the participants were thinking about the standards, the assessment, and their instruction at the topical level. This finding was further developed through a topical presentation of the alignment data. Part of the focus group discussion involved looking at the range of knowledge correspondence results from a topical perspective. Instead of looking at what percentage of the objective were assessed, the participants examined data to see what topics were assessed and what topics were missing in the assessment. Looking at a topical view of alignment data is also an approach that is used in the SEC methodology (Porter & Smithson, 2002) as a way to create a common language for comparisons among standards, assessment, and instruction. Presenting the results to the participants grouped by topics built on the work the participants did with the table view of the Math ABE standards where the objectives were grouped by topic. Additionally, in their

answers to the debriefing questions, the participants seemed to focus on the big ideas (place value, addition/subtraction, etc.) and not the specific objectives within a strand.

Reporting Webb' s range of knowledge correspondence variable but at the topical level allowed the participants to see what proportion of the topics were actually covered and served as better guidance as to what topics should be addressed in future years of test development. The information in Table 31 shows the range calculation at the topical level. Strands with 40%-50% coverage are shaded as weak and strands with less than 40% are shaded as low. The information in Table 31 includes the specific topics the participants noted were absent in their survey debriefing questions. The participants' quick reactions captured through the survey debriefing questions after matching items to objectives support the findings of the range calculation at the topical level.

Table 31 – Range Results at the Topic Level

Level	Strand	# Topics	Average topic coverage	Topics noted as Absent by the Participants in the Debriefing Questions
2	N	9	70%	
	P	11	39%	Number lines, inequalities, and missing variables
	S	6	54%	
	G	4	56%	
3	N	14	43%	Rounding and order of operations
	P	6	75%	
	S	6	56%	
	G	11	39%	Symmetry, triangles/angles, and measurement
4	N	16	48%	Percentages and fractions
	P	7	64%	
	S	6	50%	
	G	10	47%	Circles
5	N	11	47%	Number Sense items and fraction/proportion
	P	5	70%	
	S	5	80%	
	G	8	46%	Angles/triangles, area/perimeter/volume, and symmetry

Key:

	=weak (40%-50%)
	=low (under 40%)

The topical view of the data was also presented to the participants during the focus group to show, on average, the number of item-objective matches per topic. These results applied Webb's approach of counting all item-objective matches and then averaging the results across all of the participants. Appendix I details these results per level and the information in Table 32 provides an abbreviated sample from level 2. Each table lists for each level, the topic, the item number that was coded to that topic by at least one participant, the total number of participants coding that item to that topic, and then that average number of item-topic matches for that topic. Looking at the data

this way shows which topics were heavily emphasized, which topics had only a weak representation (for example, less than 1 average item-topic match), and which topics were not addressed at all (Missing topics).

Table 32 – Sample of a Topical View of the Items within each Level

Level 2 - Missing topics

Number Sense	Calculator usage
Patterns, Relations, and Algebra	Missing variables
	Number line
	Place value/inequality
Geometry and Measurement	Area/perimeter/volume
	Shapes-properties

Level 2 - Assessed Topics

Path	Level 2			
Count of Item				
Topic	Item	# Hits	Average Hits	New Topics
Addition	10	1		
	31	1		
	32	2		
Addition Total		4	0.67	
Decimal/ Percentage/ Fraction - Equivalent	13	1		
	39	1		
Dec/ Perc/ Frac - Equivalent Total		2	0.33	Added
Decimals	9	2		
	28	4		
Decimals Total		6	1.00	Added
Division	3	1		
	6	1		
	10	5		
	13	1		
Division Total		8	1.33	
Fractions	1	6		
	6	5		
	39	1		
Fractions Total		12	2.00	
Map skills/ Coordinates	17	6		
	34	6		
	35	6		
Map skills/ Coordinates Total		18	3.00	

In this sample from level 2, there was unanimous agreement that there are 3 items related to the topic Map Skills/Coordinates. However, there is only an average of 0.67 items related to the topic Addition. This analysis also lists which topics have no items at all (Missing Topics). A number of these topics also related to areas of weakness listed by the participants in their debriefing comments. This analysis shows topics that may receive minimal to no attention in the current version of the test and may need to be addressed through future pilot items.

Finally, this analysis shows which topics did not have objectives at the level of the assessment but still had items matched to an objective associated with that topic from another level (Added column). For example, decimals is not addressed in the level 2 curriculum frameworks but participants did match level 2 items to an objective associated with decimals but from another level in the curriculum frameworks. These results can also inform the framework revision process as maybe there should be an objective addressing decimals at level 2 given that students are able to answer questions related to this topic.

During the focus group discussion the participants had a higher level of interaction with this topical analysis than with any of the earlier results presented as part of the Webb methodology. The participants thought the topical analysis was very helpful, related to their approach to instruction, and was a meaningful way to represent the assessment. The participants really appreciated this view of the data and found it very meaningful. Judy stated, "I think we teach by topic, not by individual objectives. I think you want to cover main topics as well as objectives. I think that will be more valuable." And Sabrina supported this point by focusing on how the students will view

the test. She stated, "I think as the students are testing the topic is what they see. So they' ll come back and say the whole test was clock questions." Looking at this analysis the participants had more observations. They noted that on the level 2 test, having 3 questions out of 40 on Map skills/Coordinates was excessive. An observation such as this was not possible with the Webb results that were at the objective level or the standard level. Judy also noted that at level 2, "looking at mult/div I see 0.33 questions. This is what we were saying before." This analysis supported the participants' thoughts that the Number Sense was more heavily weighted to addition/subtraction and did not deal enough with multiplication/division and when to do each operation.

The topical results showed what topics were missing or over represented. Judy agreed that the topic view helped to really show where the emphasis was placed within each strand. She stated, "It goes back to the topics again. That' s what makes the most sense to me. You have some questions on this and some on this so you don' t end up with a lot on one." The missing topics supported many of the findings the participants' put down in their debriefing comments. Beyond the missing topics, anything less than 1 is barely being covered so the participants agreed those topics should be looked at for future pilot testing efforts. After reviewing topics that had minimal coverage, Mary asked, "Is there something that is terribly heavy?" and then the participants discussed this question using the topical results. The participants really worked with the data and thought about whether a certain topic should be highly represented or not.

The participants thought that analyzing the data at the topical level provided valuable information for future assessment modifications. The topical data showed more specifically about areas that were over or underemphasized and supported the

participants' conclusions after the item-objective matching process. For example, Judy stated, "In thinking about how we said there was too much stats but it didn' t show too much stats, what I' m thinking in my head is there was too much mean. That brings us back to topic. There is way more to stats than finding the average." This comment demonstrated that while the assessment may meet the percentages set forth in the test specification table, it is not necessarily covering the range of topics within that strand.

The range of knowledge correspondence calculation within the Webb methodology, however, did not help to inform what needs to be addressed because of the broad number of objectives within this strand. Looking at the topical data illustrates what aspects of Statistics, or any strand, are emphasized and allows the participants and test developers to modify these in a more meaningful way. The participants had a strong sense of what was over or underrepresented in the assessment based on their item-objective matching and this was supported in the results viewed at the topical level.

Difficulty viewing the assessment as a whole. At the conclusion of the item-objective matching process, during the debriefing questionnaire, the participants were asked to think about what was over or underrepresented in the assessment and how well that test level was aligned to the curriculum frameworks. The participants found these questions difficult to answer. During the item-objective matching process the participants found that they were very focused on looking at each item individually and there was no mechanism in place to see a summary of the objectives that had been matched. Given the narrow view and lack of a summary format, participants found it difficult to step back and think about the assessment as a whole. Judy stated, "When

you asked what was missing, I felt like I was paying so much attention to the individual things I wasn' t looking at what was missing. And I would have liked to have been able to do that better. I couldn' t look at the test as a whole." Even though the participants had been shown the summary questions in advance and answered them after each level, without a system in place to help the participants move from the individual to the whole, any view of the assessment as a whole was very challenging.

Mary did use her own method to check off objectives as items were matched to them. She stated, " When I choose an objective to match an item I highlighted the objective. So I started to get a sense of what objectives within a certain level I thought were being hit and then the ones that weren' t being touched." This was an interesting method and she suggested it could be built on in the future to incorporate a more thorough system to help the participants to think about the assessment as a whole.

The participants were more confident writing about what was over or underrepresented than in think about how well aligned the assessment was to the curriculum. Mary stated, " How would I know if it was acceptable or not? It was just a feeling?... I didn' t know what acceptable should be." Judy supported this point and related it to her inability to step back and see the test as a whole. She stated, " And because I couldn' t see the overview I couldn' t really say...I felt I wasn' t qualified to give it. I didn' t look at the questions overall and I didn' t have enough background to say whether it was acceptable or not." The participants' summary view was shown in Table 28, but the comments here demonstrated the lack of confidence the participants had in their rating. The goal of this summative debriefing question was to succinctly understand the participants' general feedback about the degree of alignment but it may

not have been a fair question to ask without more supports put in place for the participants to facilitate making a conclusion.

Effect on Instruction

The participants' involvement in the alignment process influenced the way they thought about their instruction. There were three themes within this area of the results. First, for some participants the alignment process seemed to be more limiting than the way they wanted to approach their instruction, while for others they saw many benefits to their instruction through their participation in the alignment process. Second, the participants appreciated looking at the cognitive dimensions of the items and the objectives and thinking about how they can use that in their classrooms. Third, the participants gained an increased knowledge of the frameworks that they can now integrate into their lessons. Each of these themes will now be discussed.

Limiting and supporting influences. The participants had different thoughts about how their involvement in the alignment process would affect their approach to instruction. Two participants stated that they did not want it to influence their approach to instruction because matching items to objectives was too stifling an approach then thinking about the math required in a problem. Beth stated,

> "We' re a bit at cross purposes. We' re [Mary and I] part of this math initiative and one of the things we are doing there is there is we have a problem and different people show how they solved it. And there are some classic problems that are solved different ways and we think that is wonderful. And this [alignment] process is exactly the opposite. This [process] is 'define what you are doing.' I personally lean towards the former, not this, so I' m not sure this is going to affect my teaching. I' m not sure that I want it to."

For Beth, the idea of forcing each item to match to objectives seemed to constrain the way she wanted to think about the math in a problem.

Judy understood Beth' s point but interpreted the influence on her teaching in a different way. She stated,

> "I agree. I always tell my students that they are not going to come and check your scrap paper to see how you solved it. But to me, when I look at this [alignment process] I don' t see it as saying you*have* to do it this way. I see it as saying [to me], ' Make sure you do this. Make sure you cover it.' I don' t see it as saying *how* you have to do it. And that I think in doing this whole process has brought all of this to life. To make sure I am covering this and this."

The alignment process helped Judy to see what she needed to cover through her instruction but not necessarily how that concept needed to be taught.

Furthermore, Judy saw the benefits to understanding the assessment creation process in terms of how she worked with her students. She stated,

> "Ithought the whole process was interesting. I never spent a lot of time looking at tests and individual items and how they fit together. It was an interesting thing to me to see how that works. I feel like now when I give the tests I can understand what I' m giving them and say to students this kind of question is on here because you need to be able to do this or these are the different areas and they' re all important."

Thinking about the objectives, the test specifications, and how the assessments fit with the standards helped Judy to think about making the assessment more meaningful for her students.

The participants also learned about how to present information to their students in different ways. Melissa discussed the importance of the language that is used, " being specific and being clear." The interaction between the frameworks and the assessment helped the participants to see the " language of math" and how this can be taught in different ways. Judy also found the alignment process helpful to think about how to teach concepts in different ways. She stated, " And also for me to think about how else

could I present this kind of item. It really did make me think about that. That this is something they are going to mess up so how are we going to approach this, what other ways can we think about this, or how can I write other questions like this for them in class." Understanding the frameworks, how they are operationalized in items, and the different cognitive levels of application helped the participants to think about how to approach the instruction in their classrooms.

Application of cognitive level understanding. The difference in cognitive levels was also an area that was reinforced to the participants. Sabrina stated, " For me the cognitive levels was the most informative. I never thought of that. It was, ' Can you do it - fine.' It was just a check off and move on." As a newer math teacher it was important for Sabrina to learn more about the cognitive dimension. Melissa also supported this point and stated, " It will also make us think of that problem at the different cognitive levels."

The participants saw how items could be modified to fit different cognitive levels and to think about the distinctions between the cognitive levels. Melissa also supported this point and stated, " I think it will make you refocus. When you are presenting the material you present at Knowledge and Comprehension but then you want to bring it to the next step where you are applying it. Maybe you actually want to bring it to the third step with critical thinking." Knowledge of the cognitive levels will inform the way the participants think about presenting information to their students to continue to challenge their students' understanding.

Integrating new knowledge of the standards. The participants also benefited from the in-depth analysis of the standards. Sabrina, a less experienced math teacher,

found the interaction with the standards very helpful. She stated, " Well I certainly know the frameworks a lot better..It will help with planning. I mostly do tutoring for the GED and I didn' t have much of an understanding of how it works and how the frameworks work. So it helps me to know where I can go back to." She will now be able to build on what she learned as she develops lesson plans for her tutoring.

The participants benefited from seeing individual items and thinking about why students might have struggled with the item. This often involved thinking about what the specific objective was asking of the student. For example, one item seemed quite easy to the participants but many students struggled with it on the pilot testing. The item had to do with the names of coins. The participants discussed the importance of teaching students the names of the coins in the context of understanding what the coins are worth. The participants agreed that the math behind this problem was not challenging but the language of math, the coin names, needed to be better taught. In this way the item review process helped the participants to think about what they do in their classroom.

CHAPTER 5

DISCUSSION

Using the Webb methodology, the MAPT for Math appears well aligned to the Math ABE standards. The MAPT for Math met Webb' s alignment criteria for depth-of-knowledge consistency and balance of representation across the standards in all four levels of the assessment. The assessment failed to meet the criteria for categorical concurrence for three out of sixteen standards and specific recommendations for additional items were noted in the results. The weakest area of alignment for the MAPT for Math was the range of knowledge covered by the assessment. Only one standard met the criteria for this dimension. The source of challenge, general comments, and debriefing questions offered more specific recommendations to improve the MAPT for Math. Many of the debriefing comments noted concerns about topics that were not covered enough or cognitive areas that may be underrepresented. These comments illustrated that some aspects of the alignment, as viewed from a teachers' perspective, might be masked by the Webb methodology approach.

The second research questions explored how teachers' participation in the alignment process influenced their view of the standards, the assessment, and their approach to instruction. The results for this question evolved from the discussion and observations throughout the alignment process, responses to open response questions, and the focus group discussion. There were four main areas in the results for the second research question. Within each area specific themes were discussed. First, the participants found the opportunities for discussion very beneficial and wanted more opportunities to share their ideas and questions. The discussions augmented the process

of reaching consensus where necessary, facilitated a common understanding of the terms used in the alignment process, and provided the participants with a deeper understanding of the alignment process. Second, involvement in the alignment process influenced the participants' view of the standards. Participants gained a deeper understanding of the Math ABE standards by seeing the standards in a table format and thinking about ways the standards could be modified in the future. Third, involvement in the alignment process influenced the participants' view of the assessment. The participants gained a greater appreciation for test construction, appreciated viewing the results of the analysis in a topical framework, and had difficulty stepping back and viewing the assessment as a whole immediately following the item-objective coding process. Fourth, involvement in the alignment process influenced the participants' thoughts about instruction. The participants believed the alignment process could have both limiting and supporting influences on their approach to instruction, but agreed the increased knowledge of cognitive levels and curriculum frameworks will help to guide their instruction.

The Webb alignment methodology is a widely used approach to demonstrate that an assessment measures the content and cognitive expectations as expressed in the state standards. The No Child Left Behind legislation requires states to demonstrate that their statewide assessment aligns with the statewide curriculum frameworks. A survey by Martone, Sireci and Delton (2007) contacted 24 Chief State School Officers to find out what method was used for state test-state curriculum alignment. Seventy-nine percent of the responding states used the Webb methodology to demonstrate state test-state curriculum alignment.

While this method is used in many states, a number of concerns about the Webb methodology have come into focus through this research study. An analysis of the results to this study' s two research questions leads to implications in terms of changes for the Webb alignment methodology as well as improvements for the MAPT for Math and the alignment review process in general. First, the lack of a step to measure how well the assessment accomplished what the test specifications determined will be discussed. The Webb methodology focuses on the alignment of the assessment to the standards but does not examine how well the assessment accomplishes what it was designed to do. The lack of a step to confirm the original test specifications of an assessment is a significant omission that will be discussed in greater detail below. Second, concerns and suggestions for each of the alignment dimensions used in the Webb methodology will be addressed. Third, suggested modifications to the MAPT for Math are listed based on the results of the alignment process. Fourth, the remainder of the discussion will review some improvements to the alignment process in general. The improvements include: earlier and more in-depth exposure to the state standards, improved training, increased discussion, flexibility in the implementation, and a new institute design for the alignment process.

Alignment to the Test Specifications

While the Webb methodology attempts to ensure that the assessment adequately represents the breadth and depth of the standards, it is also important to ensure that the assessment accomplishes what it was designed to do. As discussed in the literature review, a key component of test development is the determination of the test specifications. The Webb methodology does not include a step to confirm or match the

alignment results to the test specification table that guided the test development process. Without a comparison to the test specifications document, every strand is treated equally and assessed at the same level. This lack of comparison in the Webb approach is a weakness since it is in the test specifications that the relative emphasis is determined for each strand and cognitive level across the assessment levels.

The participants in this study appreciated comparing Webb' scategorical concurrence results to the requirements set forth in the MAPT for Math test specification table. Through this comparison participants saw which strands and cognitive levels were meant to be emphasized, how well goals were met, and if those goals should be revised based on the forty item distribution they saw. Table 30 illustrated the benefits of requiring a level of participant agreement and applying the results to the target levels set in the test specification table. It will be important to include a step in the Webb methodology going forward to enforce a level of participant agreement and ensure that the results for categorical concurrence and depth-of-knowledge consistency are meeting the specifications of the test design process.

The comparison of the minimum participant agreement to the test specifications also showed that the percentage of split agreement for the cognitive levels is quite high. Improved training regarding the distinctions among the cognitive levels should increase the level of agreement. Improved cognitive level training might help to place some split items in Analysis, Synthesis, and Evaluation, however, future pilot testing efforts should continue to target Analysis, Synthesis, and Evaluation items. These conclusions represent the valuable information that can be gained from comparing the data to the

original test specification table while also setting a minimum level of participant agreement.

Dimension Concerns

The Webb methodology outlined the dimensions used to analyze the degree of alignment and the rational behind the criteria for acceptable alignment. Through the application of this methodology, some concerns with the dimensions evolved. The concerns and possible modifications are discussed below.

Categorical Concurrence

In the Webb methodology, categorical concurrence was the dimension that examined if the assessment and standards measured the same content. There were two concerns with the calculation of categorical concurrence. First, there was no required minimum level of agreement among the participants about what the item was truly measuring. The Webb methodology is based on average hits, where a hit is any item/objective match, regardless of how many participants agree with a participant about what that item measures. Participants could match an item to up to three different objectives and each is given equal weight. Thus if the six participants were split among the four strands in terms of what an item is measuring, each of their item/objective hits would still be included in the categorical concurrence calculation.

Item 23 on the Level 4 assessment illustrated how this average methodology masked potential trouble in an item because each hit was equally counted. This item was matched to 4S by four participants, 4G by 1 participant, 5P by 1 participant, and 5G by 2 participants (items can be matched to more than one objective and can be matched outside of the intended level). Thus there was less than 70% agreement about what this

item was measuring at a basic strand level yet this was not highlighted in the Webb methodology. While categorical concurrence was calculated for each participant individually and then averaged across participants to balance out any extremes, there was still no criterion for an acceptable level of agreement about what that item was truly measuring. It would be more appropriate to ensure that there is some level of agreement about what the items are measuring using the guidance provided in traditional content validity.

The second concern with the calculation of categorical concurrence was the use of six items to measure this dimension. The statement of " 6 items" wasmisleading because what were actually measured were hits or item/objective matches. The hits corresponded to how many objectives were assessed by an item, regardless if that was a unique item. For example, it was possible that one participant had 6 hits for the standard 4 Number Sense. Those six hits could have been one item matched to three objectives within Number Sense and a second item also matched to three objectives within Number Sense giving only 2 unique items. While this is an extreme example, this issue did arise with the participant data in this study.

For example, Len had 8 hits (item/objective matches) for the standard 4 Statistics and Probability. These hits were only 6 items, with 2 items matched to more than one objective in Statistics and Probability. The average number of hits for standard 4 Statistics and Probability was 7.67 based on the Webb methodology. If only unique items were included in the calculation the result is an average of 6.83 items for this standard. This number truly represents the average number of items each person viewed as measuring objectives within 4 Statistics and Probability. This result did not

change the conclusion that this standard did meet the requirements for acceptable categorical concurrence. Including unique items, however, is a more accurate representation of the Webb requirement for 6 items per standard and might have an impact on whether other standards meet the requirements for categorical concurrence.

Depth-of-Knowledge Consistency

Depth-of-knowledge consistency examined how well the items measured the cognitive complexity expected from the objectives to which the items were matched. This dimension was also calculated for each participant and then was averaged across participants. There were two concerns with the calculation of depth-of-knowledge consistency. First, the calculation did not take into account the range or balance dimensions. The range of knowledge correspondence dimension stated how many of the objectives within a standard were measured. The calculation for depth-of-knowledge consistency did not factor in how many objective/item matches were within that standard in all. For example, if there were only one objective/item match and it is above, that standard would meet the depth-of-knowledge criteria. It would be helpful to understand the range of knowledge correspondence results and then examine the depth-of-knowledge consistency.

Webb noted the interplay between balance of representation, how evenly the objectives are assessed, and the depth-of-knowledge dimension. The Web Alignment Tool manual (Webb, Alt, Ely, & Vesperman, 2005) notes that the depth-of-knowledge calculation assumes the assessment is balanced and the items are not clustered around a few objectives. Table 33 shows an example of three objectives assessed through six items and the level of cognitive match for each (Webb, Alt, Ely, & Vesperman, 2005).

Table 33 – Example of Item-Objective Matches and Cognitive Level Classifications

Item	Objective the Item is Matched To	Cognitive Level of the Item Compared to That Objective
1	2N-1.1	Below
2	2N-1.1	Below
3	2N-1.1	Below
4	2N-1.1	Below
5	2N-2.1	At
6	2N-3.1	At

In this small example, the depth-of-knowledge would be 66% since two-thirds of the objectives (objective 2N-2.1 and objective 2N-3.1) were measured by items that were at or above the level of that objective. Only objective 2N-1.1 was measured by items that were below the objective. In reality, a student would not see items that were at the correct cognitive level item because the standard was not balanced.

The second concern is that the depth-of-knowledge consistency dimension analyzed how many item matches were at or above the cognitive expectations of the matched objective. This helped to ensure that the assessment was not dumbing down the curriculum. However, this result did not state if the items for each objective were actually at the cognitive level specified by the objective. The depth-of-knowledge results do not consider the different cognitive level in isolation so it is only reported that for 2 Number Sense, for example, 56% of the assessed objectives were assessed by items that were at the cognitive level of the objective and 32% were above. This result does not include whether all of the 56% objectives were Knowledge and Comprehension objectives specifically.

While 50% of the objectives in the Math ABE Standards were coded as Knowledge and Comprehension at level 2, an examination of the Webb alignment data

showed that the participants thought that only 28% of the items were at this cognitive level (based on a minimum agreement of 4 out of 6 regarding the cognitive level of the item as reflected in Table 30). This was a complaint that was voiced in the focus group discussion where participants thought level 2 needed more rote calculations to test a basic understanding. The depth-of-knowledge consistency dimension in the Webb methodology did not capture the different expectations at the specific cognitive levels. In this way the results did not accurately show what the standards or test specifications listed as the desired content.

Range of Knowledge Correspondence

Range of knowledge correspondence examined what percentage of objectives within a standard was measured. To meet this criterion at least 50% of the objectives within a standard needed to be assessed. This dimension was calculated by determining how many objectives were assessed by at least one item for each participant. Then the average percent of objectives was calculated to determine the average range of knowledge that was assessed. Similar to categorical concurrence, this calculation also did not take into account the level of agreement among participants so each hit was equally counted.

There were two concerns with the calculation of range of knowledge correspondence. First, the Webb methodology stated that if an item was matched to more than one objective it should fully measure each of those objectives. When the participants discussed this calculation they wondered if they should have matched more items to more than one objective. The participants thought if they had spent more time they could have found more matches for each of the items. However, every item-

objective match assumes the item fully assesses all the skills encompassed in that objective. If participants had matched each item to more objectives that might have helped meet the range criteria but each item might not fully measure each objective. It would be important to include a way to record if an item only partially measures an objective as used in the Achieve methodology (Rothman et al., 2002).

Second, the Math ABE standards are penalized in a way for being very detailed. Some standards had up to 30 objectives making it difficult to meet the 50% requirement in a 40 item assessment. For example, there are specific objectives for each of the four operations. The Math ABE standards could be revised to combine objectives around more general ideas. A reduction in the number of objectives would make it easier to meet the range of knowledge criteria. But again each item might only measure a portion of the more generally stated objectives. In the Webb methodology there is no way to judge the quality of the item-objective match. In the future it would be helpful to include a code to signify when only a part of an objective is met by an item or when the objective is too general to be truly assessed. This type of analysis is included in the Achieve approach (Rothman, 2003; Rothman et al., 2002).

Balance of Representation

Balance of representation examined how the items were weighted around the objectives that were assessed. To meet this criterion the index calculation had to be 0.7 or greater. Although participants were concerned about some topics being over-assessed, every standard met the requirements for acceptable balance of representation. Based on sample hypothetical calculations it seemed difficult for a standard to be unbalanced. Examples of different sample calculations for the balance index are listed

in Table 34. Webb' s balance of representation index was used

$$1-(\sum_{k=1}^{0}|1/(O)-I_k/(H)|)/2$$

where Obj (O) shows the number of objectives that are assessed, hits (H) shows the number of item/objective matches, Item (I) lists the number of items matched to that specific objective (k). The Absolute Value column (Abs) shows the calculation that is within the absolute value portion of the balance index for each objective. This calculation is then divided by two and subtracted from one.

Table 34 – Hypothetical Balance Calculations

Perfect		
	Obj	Hits
	8	8
Obj	Item	Abs
4N-1.1	1	0
4N-1.10	0	
4N-1.2	0	
4N-2.1	1	0
4N-2.2	1	0
4N-2.4	1	0
4N-2.5	1	0
4N-3.1	0	
4N-3.10	0	
4N-3.11	0	
4N-3.3	0	
4N-3.4	1	0
4N-3.5	1	0
4N-3.6	1	0
4N-3.7	0	
4N-3.8	0	
Total	8	0
		0
Index		1

Perfect		
	Obj	Hits
	1	8
Obj	Item	Abs
4N-1.1	8	0
4N-1.10	0	
4N-1.2	0	
4N-2.1	0	
4N-2.2	0	
4N-2.4	0	
4N-2.5	0	
4N-3.1	0	
4N-3.10	0	
4N-3.11	0	
4N-3.3	0	
4N-3.4	0	
4N-3.5	0	
4N-3.6	0	
4N-3.7	0	
4N-3.8	0	
Total	8	0
		0
Index		1

Accept		
	Obj	Hits
	16	22
Obj	Item	Abs
4N-1.1	7	0.26
4N-1.10	1	0.02
4N-1.2	1	0.02
4N-2.1	1	0.02
4N-2.2	1	0.02
4N-2.4	1	0.02
4N-2.5	1	0.02
4N-3.1	1	0.02
4N-3.10	1	0.02
4N-3.11	1	0.02
4N-3.3	1	0.02
4N-3.4	1	0.02
4N-3.5	1	0.02
4N-3.6	1	0.02
4N-3.7	1	0.02
4N-3.8	1	0.02
Total	22	0.51
		0.26
Index		0.74

Poor		
	Obj	Hits
	7	13
Obj	Item	Abs
4N-1.1	7	0.40
4N-1.10	1	0.07
4N-1.2	1	0.07
4N-2.1	1	0.07
4N-2.2	1	0.07
4N-2.4	1	0.07
4N-2.5	1	0.07
4N-3.1	0	
4N-3.10	0	
4N-3.11	0	
4N-3.3	0	
4N-3.4	0	
4N-3.5	0	
4N-3.6	0	
4N-3.7	0	
4N-3.8	0	
Total	13	0.79
		0.40
Index		0.60

Only in the last instance (Poor), where 7 objectives were assessed but 7 of the 13 items were matched to only one objective, was the balance index less than .7. When these scenarios were shown to the participants in the focus group they agreed that even

the .74 example would be unacceptable. Participants discussed used the example of the 4 time questions at level 2 to further illustrate their concern with this dimension. That standard (2 Geometry and Measurement) still met the criteria for acceptable Balance of Representation but all of the participants agreed 4 questions related to time were too many.

A possible modification of the balance of representation dimension would be to repeat this analysis at the topical level. The formula would be the same but instead of looking at specific objectives, topics would be used. Using the same variables as Webb' s balance calculation,the variables would now be defined as:

O – Number of topics that person matched items to

H – Total number of hits (item/topic matches) that person had for that standard

I – for each topic k, number of items matched to that topic k

In this instance the 4 time questions would be attached to the topic of time, not split across 2 objectives. This type of analysis would better capture the way the participants thought about the assessment in terms of what topics were over or underrepresented. However, it would be important to revisit the topic classifications to determine if there is still too fine a demarcation between topics. For example, it could be better to combine Addition, Subtraction, and Addition/Subtraction topics into one topic. After ABE practitioners agreed about the proper topic categorization the balance calculation could be replicated to determine if specific topics are given more weight in the assessment than the participants think is representative.

Improvements to the MAPT for Math

Based on the results applying the Webb methodology, the participants' discussions and feedback, and additional analyses and results at the topical level, some modifications to the MAPT for Math can be suggested. The Webb methodology results illustrated that some specific items should be added to specific strands to meet the requirements for categorical concurrence. The categorical concurrence criteria could be met if the next version of the MAPT for Math included two additional Patterns, Relations, and Algebra items at level 2 and two at level 3. These items could replace Number Sense items which had the greatest number of hits at each level. An additional three Geometry and Measurement items could also be added to level 5. These items could replace three Statistics and Probability items which had the greatest number of hits.

The range of knowledge correspondence dimension results showed that the MAPT for Math must ensure that the assessed objectives rotate so the entire breadth of each standard will be assessed over a multiple year assessment period. If this did not happen, the MAPT for Math could be responsible for a narrowing of the curriculum since only a subset of the objectives would be consistently assessed. Replicating this alignment study over the course of a three-year period would help to ensure that the curriculum is not being narrowed through what is assessed. Results of the topical analysis and the participants' feedback helped to inform specific topics that should be assessed in future versions of the MAPT for Math and these results should drive item selection for the unassessed objectives.

The cognitive dimensions should also be reassessed. All assessment levels require a greater number of higher cognitive level items. This might best be met with more innovative, open-response type items where students can better demonstrate higher level thinking skills. Finally, the calculator placement should be revisited to determine if it is always better to place the calculator to the right of the problem rather than between the question and the answer choices. Specific recommendations for each level are presented next.

Level 2

- Review source of challenge items that require possible modifications to see if they need to be adjusted and re-piloted or replaced.
- Add two Patterns, Relations, and Algebra items that are clearly tied to an objective within this strand.
- Replace two or three time related questions.
- Pilot items should include more: basic operations (especially multiplication/division), area/perimeter/volume, calculator usage, symmetry, shapes, number lines, place value/inequalities, and missing variables.
- Include more Knowledge and Comprehension items in place of Application items.

Level 3

- Review source of challenge items that require possible modifications to see if they need to be adjusted and re-piloted or replaced.
- Add two Patterns, Relations, and Algebra items that are clearly tied to an objective within this strand.

- Pilot items should include more: exponents/roots, manipulating fractions, symmetry, triangles/angles, mean/median, rounding, solving expressions, measurement, and order of operations.
- Include more Analysis, Synthesis, and Evaluation items in place of Application items.

Level 4

- Review source of challenge items that require possible modifications to see if they need to be adjusted and re-piloted or replaced.
- Pilot items should include more: number lines, calculator usage, equivalencies between decimal/percent/fraction, division, fractions, place value-inequality, probability, ratios, shapes, measurement, and circles.
- Include more Analysis, Synthesis, and Evaluation items in place of Application items.

Level 5

- Review source of challenge items that require possible modifications to see if they need to be adjusted and re-piloted or replaced.
- Add three or four Geometry and Measurement items that are clearly tied to an objective within this strand.
- Pilot items should include more: division, manipulating fractions, integers, measurement.
- Include more Analysis, Synthesis, and Evaluation items in place of Application items.

A great deal of specific information was provided from the alignment review process to inform future assessment modifications.

Improvements to the Alignment Review Process

Throughout the alignment review process participants made suggestions about how the process could be improved. First, participants all requested time to become more familiar with the Math ABE Standards prior to the alignment review. Second, other aspects of the training could also be improved. Third, throughout the alignment process the participants could have benefited from increased discussion. Fourth, flexibly implementing the alignment process was important to ensure the needs of the participants were met. Each of these improvements will now be discussed in greater detail.

Increased Familiarity with Math ABE Standards

The alignment process began with an assumption that the ABE teachers were familiar with the Math ABE Standards. The document is only about a year old and is a 100+ page long list of different objectives within different standards. Throughout the alignment process it was apparent that the participants could have used an activity prior to the alignment process to familiarize themselves with the standards and objectives.

In the comments following the objective coding phase, Melissa stated she would have liked to have reviewed the frameworks more before the meeting. Unfortunately I did not plan an activity or task to facilitate this review process. In the comments after the item-objective matching process the participants expressed the frustration they felt with the item-objective matching process because of the difficulty of finding objectives.

Participants also commented on the specificity of the objectives making the matching process difficult.

Even with the table layout of the objectives, it was still difficult to navigate through the 20 pages. A few improvements were suggested for the table layout of the objectives. The document needs a table of contents to show what topics are on specific pages. Also, the topic column should be highlighted to show that this is new information the teachers might not be familiar with. In general, more time spent familiarizing the participants to the topic layout would have been helpful. I could have had treasure hunt type searches, questions to ask about the number of objectives within different topics, and just more general discussion about the layout.

These types of activities were used in both TIAN and the Math Professional Development Initiative so four out of six participants had had some more in-depth exposure to the frameworks. Len shared an experience from TIAN and how that could help with this process. He stated, "[In TIAN] we were given an activity and told to hunt through the standards at a certain level and to find all that applied to that activity... [You could] do that in the lab and train people to scan through the documents and use the find function." Mary also shared an experience from her Math Professional Development Initiative. She stated, "I did that at a workshop as well where we did an activity and then had to go back to the frameworks at a certain level and find what objectives it matched. I found it very helpful. It really made you focus on what you were measuring."

Melissa, who did not participate in TIAN or Mary' sprofessional development activity, noted this difference and stated,

> "It was very interesting to see the way Mary and Beth [participants in the Math PD Initiative] looked at things from the way I did to the way Judy and Len [participants in TIAN] did. I could see where they were coming from but they were so familiar with the frameworks. I mean they were like, ' Well, I know there is another one that this is better for it' . I' m still here looking through each one. They knew exactly what was in the frameworks but I was more like I think it is this one because I can' t find another one."

This comment supported the idea that increased training, possibly through partnering people with different backgrounds, could have helped everyone to become more familiar with the standards.

Increased Training

Rothman' s(2003) review of alignment approaches emphasized the need for in-depth training about the alignment methodology and this conclusion was supported in this alignment study. The use of specific examples of item-objective matches would be helpful in training participants on the item-objective matching process. Participants also needed to revisit the distinctions between the cognitive levels. Revisiting the distinctions between cognitive levels might not have been as necessary if the tasks flowed over continuous days. Participants would then have built on the way they operationalized the cognitive levels through the objective coding process and applied the same general rules of thumb to the item rating process.

Participants also needed more training about primary versus secondary item/objective matches. They could have benefited from more examples of how the whole item needed to relate to more than one objective if it was matched to more than one objective. For example, items might be about adding or subtracting but also drawing information from a table. In this way the item connected to both objectives.

Participants were not clear about the use of more than one objective and it was uniformly applied.

Finally, participants also needed more training about the cognitive levels for the item/objective match. Mary noted a time at level 5 where, "I thought the objective was an ASE objective yet the question was written as an Application question. It didn' t really seem to match that way. I think that is what I found difficult." In practice, the items should be matched to the skills required by the objective and then a separate step is to look at the cognitive complexity of the item. If the objective was written as an Analysis, Synthesis, and Evaluation level but the item is only asking Application skills, this would be highlighted in the depth-of-knowledge analysis. Mary' s comment illustrated that the distinction between objective and item cognitive levels needed to be clarified.

Increased Discussion

From the objective coding process the importance of discussions among participants was very apparent. This phase of the alignment study illustrated the value of the discussion and leads me to think that more discussion in the item-objective matching process would help improve the quality of the matches as well as serve as a better means for the teachers to learn from each other. With more discussion opportunities the participants would be forced to talk through what the items are really measuring and dig deeper into the objectives to support their points. Participants would learn from each other as they listened and explained their points. Reaching consensus about item-objective matches would take more time but it would be a valuable learning experience to increase the professional development aspect of the alignment process.

The consensus approach to item-objective matching was used in a Webb methodology in both Idaho (Leffler et al., 2003) and Montana (Leffler, Carr, Griffin, & Gates, 2005). Each of these applications used only three participants but I think the benefits would outweigh the increased time it would take with more people.

The discussion throughout the alignment process also helped the participants to better define the cognitive levels. As noted in the results, Len expressed strong opinions about how he thought the cognitive levels should be grouped. Through discussions during the objective coding process, the group helped this participant to see the distinction in the way the levels were currently defined and why we needed to maintain those classifications. Len was still adamant that the separation of levels was not what he would like to see and I think this influenced his rating of items during the item-objective matching process. Without the discussion during the item-objective matching phase he was more apt to rate items others saw as Application as Knowledge/Comprehension.

Feedback to the participants was very important in terms of validating their understanding. During the objective coding process, one participant in particular was grading how many times she had to change her rating as she tried to get a "perfect score". This became a bit of a running joke through the discussion process but other participants seemed to also enjoy keeping track of this statistic. Participants really seemed to want to be in the majority opinion and know how they were doing. This type of feedback could be built on with increased discussion in the item-objective matching process.

Flexibility in Approach

While each step of the alignment review process was planned in advance, it was important to be flexible in how the stages were implemented. For both the objective coding process and the item-objective matching process the steps of implementation were altered based on the needs of the participants.

The original plan for the training of the objective coding process called for the participants to rate every 10th objective and discuss their ratings. Participants asked to focus on a specific strand within a level so they could begin to develop a common understanding around a common body of knowledge. We decided this could be very helpful and went with this approach to training instead. The training started with level 3 Pattern, Relations, and Algebra. There are 14 objectives in this strand. Participants rated the cognitive level required for each of the 14 objectives and these were discussed in depth. Participants then completed the three other strands for level 3 and these were discussed in depth. This more in-depth training process allowed the participants to discuss the "lessons learned" andthe "rules of thumb" developed through this process were then applied to the rating process for the other levels.

Through this initial discussion of level 3 the participants found it helpful to focus on the verb in the objective to see what the objective is asking of the student. Objectives starting with "State" or "Count", "Read", and "Compute" were often seen by the participants as Knowledge and Comprehension skills. Objectives starting with "Find", "Make", "Show", and "Convert" were often seen as Application skills. Then objectives starting with "Investigate", "Use", "Extend", and "Choose" were often seen as Analysis, Synthesis and Evaluation type skills. However, there was not a hard and

fast rule about these terms. The participants liked the flexibility in the day so the rating process was done in a way that helped them. By reviewing a standard, then a complete level, then the remaining levels the participants gained confidence in their approach to the rating process. They also appreciated that there was humor and camaraderie among the group as they practiced listening to all participants. As one person stated, "[Drey] graciously and patiently reminded each of us to respect all voices in the dialogue while allowing good humor to continue."

Through the Web Alignment Tool (Webb, Alt, Ely, & Vesperman, 2005) the objective coding process could have been accomplished on-line. For this exercise, however, the participants believed it was very helpful to do the rating with paper and pencil rather than on a computer. Participants chose to lay out their rating sheets out so they could reference how they rated similar items at other levels. Participants also liked to see how they changed their ratings and made notations as they went along.

The item-objective matching process also was modified based on the participants' needs. The matching prœess started in a whole group setting where the participants worked with paper based versions of the Math ABE Standards. As the participants progressed in the rating process they all felt that flipping through the 20 pages of the table view of the standards was a bit overwhelming. Participants did the first 2 levels with only paper versions but then completed the last two levels at home using the electronic format of the standards. A number of comments from the item-objective matching process expressed frustration about not being able to find a specific objective related to the idea they thought the problem was asking. Using the computer to search on key words or a more thorough understanding of the layout would have

helped ease this problem. Beth stated, " When I took those 2 packets home I used the find feature. If I thought something was testing something specific I typed that word. That was so much easier. Before that, sitting here going through all of that was quite challenging." Participants all agreed that the computer helped expedite searching as they could use the Find feature in word and move through the document much faster.

It would also be helpful if the on-line tool had a feature to easily allow the participants to see what objectives they had matched items to already. This aspect of the tool would help the participants to better answer the summary debriefing question about how well the assessment aligned to the standards. Mary mentioned that she checked off objectives as they had items matched to them so she could see which ones weren' t covered or had many items matched to them. Systematizing something like this to produce a report for all of the participants would be very helpful.

Both the objective coding process and the item-objective matching process were modified based on feedback from the participants. The training approach for the objective coding process was very helpful to establish common understanding and build confidence and she be used as a guide for future alignment studies. The on-line tool should also be used for entry of item-objective matches and the electronic versions of the Math ABE Standards should be used for the item-objective matching process. These on-line features would make the objective searching process easier and expedite the analysis process.

Alignment Institute Design

Based on the comments from the participants throughout the alignment process, an alignment institute design is proposed. This design addresses the need for increased

familiarity with the frameworks, improved training, increased discussion, and the ability to flexibly implement the alignment steps. Professional development can sometimes be very difficult in the ABE field because teachers enter with diverse experiences, underdeveloped teaching skills, and no background in adult education (Belzer, Drennon, & Smith, 2001). Given the integration of standards, assessment, and instruction with a population of teachers with diverse backgrounds and levels of experience, an alignment institute type of professional development experience could help address some of the issues ABE programs face. A more formal alignment institute could also help build an example of formal coursework ABE teachers need as part of the effort to professionalize the field (Smith & Hofer, 2003).

The alignment institute would also address the specific math needs of ABE teachers. A survey of 141 Massachusetts ABE math teachers (Mullinix, 1994) found that 36% came to be math instructors "by accident" and 24% are math teachers because it is "part of the program package" of what they are required to teach. Additionally, 55% said they had no training in mathematics pedagogy. An alignment institute could help augment teachers' understanding of the standards and the assessment both in terms of content and thinking skills, and discuss potential implications for classroom practice. While this institute would not ensure that all math teachers are fully knowledgeable about all aspects of mathematics, it could provide a starting point to help teachers become more familiar with critical components of their students' math education.

The current alignment process was implemented over the course of 3 months as shown in the timeline in Figure 4. At the start of each activity the participants had to spend time reviewing past steps and recalibrating their understanding of key terms such

as cognitive level distinctions or adequate item-objective matches. A weeklong institute would help the steps to logically build on each other and the participants to gain a greater understanding of how the steps fit together. A key aspect of an alignment institute would be increased discussion about the cognitive levels and item-objective matches. The discussions would result in consensus building and agreements about how each item should be coded. But the discussions would also serve as valuable learning opportunities as the participants learned from each other about different ways to look at items and how items can be classified. Through the discussions the participants would share examples from their classrooms and build a more solid understanding of the Math ABE Standards and how they can be operationalized in assessment items.

The institute would begin with an activity that pairs participants to increase their familiarity with the Math ABE Standards. Participants would search through the Math ABE Standards to complete a treasure hunt or complete math activities and think about what objectives were used in that task. The questions or tasks would be designed to highlight some of the differences in topics across the levels. Including teachers with different backgrounds in mathematics would allow for them to be partnered to build on and extend the math content knowledge teachers bring to the meeting.

The alignment process would then start with the objective coding process. This would be similar to the way this step was enacted in the current study. First, it would start with a detailed discussion of the cognitive levels. Second, participants would code the cognitive level of the objectives for one standard (ex. 3 Number Sense) and discuss these results. Third, participants would code all of the other objectives for the other

three standards within that level. These results would also be discussed. The final step would be for the participants to code the rest of the objectives by level by strand and discuss any that were not unanimous. Based on these discussions, "rules of thumb" would develop about the distinctions between the different cognitive levels. Results of this analysis could be presented the next day to allow for a deeper understanding of the cognitive expectations across the different learning levels.

The next step of the alignment process, the item-objective matching, would immediately follow the discussion about the objective coding results. In the current alignment study this was two months later. Much of the understanding about the differences between the cognitive levels and the familiarity with the Math ABE Standards was lost in that time. Having the item-objective matching step immediately follow will ensure that the participants can continue to apply and build on the understanding they gained through the objective coding process. In this way the cognitive level distinctions should be fresh in the participants' minds. This step, however, should begin with increased training about how items can be matched to more than one objective. Specific items that illustrate this point would be used to build the participants' understanding.

Then the training would continue with additional examples to allow for the participants to have a full understanding of how to match items to objectives and how this is different and separate from the step to rate the cognitive complexity of each item. It will also be important for the participants to record when they think an item is only measuring a part of an objective. A coding system similar to the Achieve approach (Rothman et al., 2002) for Content and Performance Centrality could be used for this

aspect of the process. This increased training should alleviate some of the concerns that arose in the current study.

The item-objective matching should be done in partners and electronically so the participants can search more easily through the Math ABE Standards and enter their matches electronically. The partner discussions will help to facilitate the review process and get the participants thinking about what the item is really asking. After each level, reports showing differences in ratings could be produced to show when there were disagreements about the item-objective matches. Then these disagreements would be discussed with the whole group.

While the partner work and discussions would significantly slow down the item-objective matching process, the benefits in terms of the quality of the match and participants' learning would be great. The results of the current study illustrated how much the participants gained from the discussions and how much they wished there were more opportunities to discuss their thoughts. The discussions in the alignment institute would allow the participants to further understand what the item is asking, what objective is best measured by that item, and start thinking about what they do in their classrooms to address these issues.

The greatest limitation of the current study was in examining how the alignment process influenced the teachers' approach to instruction. Participants were not given an opportunity to really think deeply about this question and share examples of what they did or could do to address items or objectives. In the alignment institute there could be more discussion in the item-objective matching process where the participants share examples from their classrooms. This addition could make the alignment review

process have a more meaningful impact on instruction. The result would be an even closer association between the standards, the assessment, and what happens in the classroom.

The final step of the alignment process would be to present the alignment results to the participants. If the item matches are entered electronically it will be easy to create the analyses and summary tables overnight. Immediately following the completion of the item-objective matching the participants would be able to see the summary results of the alignment process. Given that all of the components have followed in quick succession, we would not have to spend time familiarizing the participants with the process, as was necessary in the current study when the presentation of the results was about a month later. With the increased discussion throughout the process, there would be even more discussion and a greater connection to the final results. The participants would feel closer to the data and more connected to the alignment process. At this stage there could then also be more discussion about how the increased knowledge of the standards and the assessment might influence instruction.

In terms of the analyses of the data, I would recommend blending the Webb methodology with features found in the Achieve and SEC approaches (Rothman, 2003). Given the increased discussions about item-objective matches, there would not be a need to set a minimum level of participant agreement. Requiring minimum levels of reviewer agreement would be an important addition to the Webb methodology if consensus was not a part of the item-objective matching process. The analysis should include a match to the test specification table as noted in the Achieve methodology.

This could be a more detailed indication of categorical concurrence and depth-of-knowledge consistency criteria currently used in Webb. Furthermore, range of knowledge and balance of representation calculations should analyze the data from the topic perspective. This would provide data that is more useful to the participants and for future test modifications.

An institute would be particularly helpful for newer teachers as they can gain an understanding of both the assessment and the standard and how these can influence instruction. Newer teachers are less familiar with all of the objectives within the Math ABE Standards but an alignment institute would help them to really understand how these objectives are applied. Beyond just teaching topics, it is important that teachers understand the different cognitive levels of thinking and how to continue to challenge students' understanding. AsSabrina noted, this knowledge of cognitive levels was not a focus in her previous approach to teaching but now she can see how this will be very helpful. Judy also supported this point and stated, "I think that too. Especially for a newer teacher you are gaining an idea of well what is an Application question. If you' re not used to presenting material you don' t know what you are presentingt as. So this gives you more familiarity. Oh this kind of question is this. I think that will be very helpful for someone that is new and hasn' t taught much of this before."

However, Beth was concerned that this type of institute could get new teachers too focused on "teaching to the test." She stated, "I' m thinking about your comment of doing this with new teachers. I don' t know as a teacher how much I want to be locked in to the assessment tool when I' m presenting material. How much do I care whether their question is going to be Application when I' m teaching. I want to teach for

Application and for higher understanding. I' m not sure how locked in I am about that." While Beth focused on specific items that we discussed in this alignment study that are on the operational test, most likely an alignment institute would use past released items so the process is more about learning about the test than teaching to the test. Having a balance of newer teachers and more experienced teachers could help as they learn from each other. While an alignment institute might be more significant for newer teachers, it is likely that experienced teachers will have their approaches validated or see areas they might need to address in the future. The benefits for teachers, in terms of learning about the standards, the assessment, and their instruction, could be an important stepping stone to developing a wider professional development initiative.

Study Limitations

There are two main limitations to this study. First, this study was conducted over a focused time period with a small population of participants. Due to the focused time period, the data do not directly assess how the alignment process influenced the participants' approach to instruction. It could have been helpful to observe and interview the teachers about their instructional methods prior to the alignment study and then immediately following the alignment process. With the current data, all that can be reported is the teachers' view of how participating in the alignment process influenced their approach to instruction.

Second, there were only six participants in this study. Given this small population it is difficult to make any reliable generalizations about the results of this study. While the participants did represent a range of experiences, it could have been

better to have an even larger number of participants to allow for more diverse viewpoints.

Further Research

One area of future research would be to look more closely at the degree of vertical alignment in the standards. Currently the objectives are mapped to topics within strands across the levels. A next step would be for participants to rate the relationship between the objectives at adjacent levels. As states evaluate growth from one level to the next, it is important to clearly understand how the skill expectations develop across the levels. Wise and Alt (2005) discuss ways to explore the relationship between adjacent objectives. The higher level objective may reflect a broader application and/or a deeper understanding. The lower level objective may be a prerequisite skill for the higher objective. Or the higher level objective may be a new skill entirely. This type of study would build on the current findings where participants noticed gaps, redundancies, and differences in terminology among the strands and levels. Wise and Alt (2005) also provide a checklist as to how to define the quality of the linkages between the objectives. The Wise and Alt paper concludes with an application of the Webb dimensions to the vertical scaling data. This type of analysis will help teachers to see how their instruction at their learning level fits within a broader understanding of instruction across the different learning levels.

Another area of future research would be to apply a more stringent requirement for participant agreement as to what an item is measured, as noted in the Herman, Webb, and Zuniga study (2005), with a larger number of participants. While the current study assessed the implication of participant agreement in terms of match to the original

test specifications, applying a requirement for participant agreement would have implications for all of the alignment dimensions. Most likely, requiring a level of participant agreement, especially at the objective level, would weaken the alignment results. However, the results would highlight the need for increased training and possible modifications to the standards and/or the assessment.

A third area for future research would be to apply the Achieve, SEC and Webb methodologies with the same participants and materials. The results from a study such as this would illustrate the comparability of findings. The participants of the study would also be able to comment on the different alignment approaches.

A final area for future research would be to further explore the influence the alignment process had on teachers' approach to instruction through detailed observations of teachers' practice. Conducting pre and post observations of teaching practice and including follow up interviews would enable a better understanding of how the lessons learned throughout the alignment process are or are not incorporated into classroom instruction.

The problem for future research to solve is how to involve teachers to continue to build the link between assessment, standards, and instruction. This study sought to understand how well an assessment was aligned to a set of state standards using an accepted approach to alignment and what the impact was on the teachers as the participants in terms of their thoughts on the standards, the assessment, and their approach to instruction. Through this study the participants gained a better understanding of the assessment and the standards and suggested modifications for both components. They also reflected on their instruction and what might change based

upon their new knowledge. Having teachers as participants also guided some of the revisions to the Webb methodology as a topical focus was brought to the data analysis which was more relevant to the way the participants approach their instruction. Building on the findings from this study will continue to augment the connection for teachers between assessment, standards, and instruction so they can move from an assessment of learning toward an assessment for learning.

APPENDIX A

COMPARISON OF THREE ALIGNMENT APPROACHES

Points of comparison	Webb	Achieve	SEC
Content	Categorical concurrence – (test blueprint) compare standards and assessments Goal: 6 items per broad content standard	Confirm test blueprint then analyze content centrality – look at degree of match Rating: 2, 1A, 1B, 0 - Able to capture standards that are too broadly written to be completely assessed	Topic coding - assessment items, standards, and instructional content are all mapped to a common content language, organized into logical groupings of topics Rating: Allows for quantitative comparisons of the instructional content emphasized in standards, assessments, and instruction

Points of comparison	Webb	Achieve	SEC
Cognitive levels	Depth-of-knowledge consistency – Cognitive demand comparison between objectives and tests Cognitive levels: recall, skill/concept, strategic thinking, extended thinking Goal: At least 50% of the items matched to an objective had to be at or above the cognitive level of that objective	Performance centrality - Cognitive demand comparison between objectives and tests but coded after determine content match Rating: 2, 1A, 1B, 0 - Able to capture standards that are too broadly written to be completely assessed Cognitive levels: focus on the verbs used in the standard and what the item is requiring – e.g. select, identify, compare, analyze, represent, use Level of challenge - captures qualitatively whether the collection of items mapped to a given standard are appropriately challenging for students in a given grade level Rating: how the overall demand compared to that expressed in the standards, if the items are skewed toward one level of difficulty, if the items are focused only on the more demanding or least demanding objectives within a standard, and if the items are skewed toward the most or least demanding part of a standard where there are compound objectives. Similar to cognitive comparison but this adds a more descriptive piece	Expectations for student performance – Cognitive demand comparison of test items, standards, and instructional focus Cognitive levels: memorize facts, perform procedures, demonstrate understanding, conjecture generalize prove, solve non-routine problems, and make connections

Points of comparison	Webb	Achieve	SEC
Breadth	Range-of-knowledge consistency – Breadth comparison between standards and assessment as judged by the number of objectives within a standard measured by at least one item Goal: At least 50% of the objectives within a standard need to be measured by at least one assessment item	Range – quantitative measure of the fraction of the objectives within a standard that are measured by at least one item Rating: Ranges between 0.5 and 0.66 are acceptable and above 0.67 is considered good coverage	NA
Distribution	Balance of representation - how evenly assessment items are distributed across objectives within a standard $1-(\sum_{i=1}^{0}\lvert 1/(O)-I_k/(H)\rvert)/2,$ where O=Total number of objectives hit for the subject domain; $I_{(k)}$ = Number of items corresponding to objective (k); and H = Total number of items hit for the subject domain Goal: Every objective assessed should be measured by at least two items	Balance – relative importance that the test items give to content and skills should be proportionately similar to what is stated in the standards Rating: Qualitatively capture which objectives within a standard seem to be over or under assessed, which items might be too much alike and therefore redundant, and how the overall set of items measures content that the participants think is important for that level	NA
Item quality	Source of challenge (added in 2005) - ensure that items are fairly constructed and are not designed to trick students Rating: Comments entered on the rating sheet	Source of challenge - ensure that items are fairly constructed and are not designed to trick students Rating: 1 or 0 if appropriate or not	NA

APPENDIX B

SAMPLE INFORMED CONSENT

Exploring the Impact of Teachers' Participation in an Assessment-Standards Alignment Study

I will participate in this study and understand that:

1) I will participate in a workshop to review and code the Massachusetts Adult Proficiency Test (MAPT) for Math and the Massachusetts Adult Basic Education Curriculum Framework For Mathematics (Math ABE Standards).

2) I will complete a pre/post survey and participate in a focus group session led by Drey Martone. The purpose of surveys and the focus group is to gather information about my opinions of the MAPT for Math and the Math ABE standards and the influence these have on my approach to instruction.

3) The alignment study will take place over two consecutive days in May. An additional one-day workshop and the focus group will be scheduled based on the participants' availability. I will be paid $600 for participation in the May meetings and an additional $400 for participation in the follow up workshop and focus group.

4) The focus group discussion will not take more than three hours and will be videotaped so that it can be reviewed and transcribed at a later date.

5) I understand that excerpts from the focus group may be included in written and oral presentations of this research. I also understand the source of the excerpts will be kept confidential and that my name, where I teach, or any other identifying information will not be used in any written or oral presentations

6) I am free to not participate in this study without consequence. I am also free to refuse to answer any questions on the surveys or in the focus group, without consequence or explanation. Additionally, I may withdraw from part or all of this study at any time.

7) I have the right to review material prior to presentation or publication. A copy of any papers or publications related to this interview will be provided to me.

8) I understand that results from this study may be included in a conference presentation and may also be included in manuscripts submitted to professional journals for publication.

9) I will be provided with a signed copy of this consent form for my records and Drey Martone will keep a signed copy for her records. If I have any further questions I can contact Drey Martone at dreymartone@educ.umass.edu.

_____ _____

Participant' s Signature Date

_____ _____

Researcher's Signature Date

APPENDIX C

COGNITIVE LEVEL/STRAND DESCRIPTIONS FOR ALIGNMENT

Massachusetts Adult Proficiency Test (MAPT) for Mathematics and Numeracy

Cognitive Level Descriptions

For Alignment Meeting, May 10, 2006

There are three cognitive levels on the MAPT for Math. Distinctions between the cognitive levels should be based on the complexity of the task required to answer the question not on the difficulty of the task. It is important to focus on type of thinking required rather than the probability a student will get the task correct.

Knowledge and Comprehension: Questions at this cognitive level test recall of information and require a rote response. These questions test a most basic understanding and ask the student to perform straight calculations. These questions are the lowest level of understanding and test students' ability to comprehend information. Questions at this level are usually not in a context.

Application: Questions at this cognitive level test skills applied to a situation. Such questions require more in-depth thinking than a rote response. In application questions, students need to make decisions about what is required to solve the question and then implement their understanding.

Analysis, Synthesis and Evaluation: Questions at this cognitive level require the most complex thinking. In these questions, students break down the information into component parts to understand the steps required, or relate different ideas together to form a common understanding. These questions might also require students to evaluate and draw conclusions based on an understanding of the components or the situation. Students might also be asked to explain or infer findings based on the context of the question or the data provided. These questions might also ask students to recommend possible approaches or solutions to a problem.

APPENDIX D

SAMPLE DEBRIEFING QUESTIONS

Summary Questions for After Each Level

For each standard, did the items cover the most important topics you expected? If not, what topics were not assessed that should have been?

For each standard, did the items cover the most important cognitive levels you expected? If not, what cognitive level was not assessed?

Was there any content you expected to be assessed, but you found no items assessing that content? What was that content?

What is your general opinion of the alignment between the standards and assessment:

Perfect alignment

Acceptable alignment

Needs slight improvement

Needs major improvement

Not aligned in any way

Other comments:

APPENDIX E

SAMPLE ASSESSMENT CODING FORM

Participant_____ Date_____
Level _____

Item #	Item Cog Level	Primary Obj	Secondary Obj	Secondary Obj	Source of Challenge	Notes
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

APPENDIX F

SOURCE OF CHALLENGE COMMENTS

Level	Item	Comment	Follow up points	Change	Type Comment
2	977	Decimals not mentioned in Level 2	No revision – item performed in the range for this level. May want to adapt the framework to introduce decimals earlier.	None	Level
2	2478	The degrees F and degrees C is not clearly labeled as Fahrenheit and Celsius - the student might not put that together.	Possible revision – it could be helpful to spell out the labels or include the label after the term in the question.	Possible	Labels
3	2377	Pictures of houses with numbers or addresses may help clarify	No revision – the important details are stated and a graphic would just add complexity.	None	Extra
3	611	None of the time is in time notation - answer should be 1:30 digital	No revision - Part of the standards is verbal expressions of time so this format is acceptable.	None	In Standards
3	1435	You can do this by adding and not using algebra = advantage?	No revision - That is acceptable. This is testing the students' understanding of the concept. Then it can be solved using different approached.	None	In Standards
3	2300	Use of word "quarters"	No revision – Quarters is an acceptable way to express one fourth.	None	In Standards
3	990	Is this just straight multiplication? Kind of confusing context.	No revision – objectives refer to scale and proportion	None	In Standards

Level	Item	Comment	Follow up points	Change	Type Comment
3	2228	Confusing - you think it's going to be perimeter by set-up context. It's got irrelevant info in it to solve and answer.	No revision – the dimensions are useful for determining the shape.	None	In Standards
3	990	Not challenging - meant to be a scale problem 4N-3.5 but numbers do not require that.	No revision – the item statistics show this to be an acceptable item.	None	Level
3	2363	I wonder why this performed at level 3? I don't know.	No revision – item statistics support this placement.	None	Level
3	285	Calculator appearing between question and answer.	Possible revision - This should be revisited by the math committee.	Possible	Calculator
3	1461	I would prefer to see the calculator under the answers. It commands attention. There is a huge separation between the questions and the answers that an LD student would have trouble with (especially when the answers aren't numbers).	Possible revision - This should be revisited by the math committee.	Possible	Calculator
3	1809	Number machine difficult concept for ESOL students (and also me! I've never heard the term)	Possible revision – objective specifically refers to In/Out table but earlier review found this term confusing.	Possible	Context
3	1809	Problem with description of "number machine" - maybe "In/Out Table"	Possible revision – objective specifically refers to In/Out table but earlier review found this term confusing.	Possible	Context

Level	Item	Comment	Follow up points	Change	Type Comment
3	1809	What's a number machine? ESOL students.	Possible revision – objective specifically refers to In/Out table but earlier review found this term confusing.	Possible	Context
3	283	Map looks more like a graph. Representational pictures rather than *s would help.	Possible revision	Possible	Graphic
3	1435	Concern about term "side by side" maybe use the word next to each other	Possible revision	Possible	Vocabulary
3	1435	ESOL students "side by side" difficult concept	Possible revision	Possible	Vocabulary
3	1449	Could be tough context/language for ESOL students (sophomores, juniors, etc.)	Possible revision	Possible	Vocabulary
3	1447	The word pictograph is unnecessarily hard and not needed. Just use chart.	Possible revision – could remove the word pictograph and replace with chart.	Possible	Vocabulary
4	297	Evaluate is operative word here?	General comment – more of a comment than a source of challenge	None	General
4	2513	Not sure actually - the level 2 seems harder than the level 3, and there is no choice for level 4.	General comment – more of a comment than a source of challenge	None	General
4	1355	This is performing at a higher level because of the names of the coins - cultural.	No revision – coin names are part of the frameworks	None	In Standards

Level	Item	Comment	Follow up points	Change	Type Comment
4	553	Performs at level 4 because they are not reading carefully.	No revision – It is important students look at what the question is asking and the reading requirement is not too overwhelming.	None	Level
4	553	This seems to be testing reading rather than math skills.	No revision – It is important students look at what the question is asking and the reading requirement is not too overwhelming.	None	Level
4	1846	Grades that 5 students... could be mistakenly quickly misread at grade 5 students - I did at least.	No revision – the change is not pertinent to what the question is asking	None	Level
4	2513	The rectangles don't really look like windows.	Possible revision – could make the shaped look more like windows to add to the context of the question	Possible	Graphic
4	848	The thermometer is difficult to see. Clear picture of full thermometer might be better.	Possible revision – could get a better picture of a thermometer although it is clearer on the screen than in the printout	Possible	Graphic
4	2366	The word "culture" is a specific science term. It might trip up ESOL students who often encounter "culture" in language classes.	Possible revision – could change the word "culture" to "sample".	Possible	Vocabulary
4	2147	The word material is confusing here. Material meaning cloth? Building material? Gender bias.	Possible revision – could change "what is the distance around the edge of the room?"	Possible	Vocabulary

Level	Item	Comment	Follow up points	Change	Type Comment
4	2147	The word "material" is a bit vague to me. Maybe just "what is the distance around the edge of the room?"	Possible revision – could change "what is the distance around the edge of the room?"	Possible	Vocabulary
5	2557	Asked to identify "equation" but answers are statements.	The question was changed to say expression not equation	Done	Done
5	2557	The standard specifically states numbers. Does that make this ASE?	General comment – this is more about the cognitive level and is not a source of challenge	None	General
5	1809	"Imaginary" NUMBER MACHINE again!	Possible revision – objective specifically refers to In/Out table but earlier review found this term confusing.	Possible	Context
5	2304	Is this cultural? Did you see if errors were by foreign born students?	Possible revision – this question related to clothes worn in certain temperatures. Different cultures may respond differently. This question might be up for replacement in the next year of testing.	Possible	Context
5	1006	Construct "has had" may be difficult for ESOL students.	Possible revision – the question could be changed to just say "had" instead of "has had".	Possible	Vocabulary

APPENDIX G

GENERAL COMMENTS

Level	Item ID	Comment	Code	Details
5	425	Where is % on this calculator?	Calculator	
2	2144	Benchmark says graph, but I think it applied here.	Framework comment	
2	1394	Is there something at a level 2 or 3 that describes multi-step arithemetic problems?	Framework comment	
2	1037	We need a "Choose the correct expression, 1 step equation, 2 step equation" standard.	Framework comment	
2	2270	2S-2.4 doesn't mention line graph, but 3S-2.5 does. Maybe "line" was left out of 2S-2.4?	Framework comment	
2	2270	Says bar graph-think it applies	Framework comment	
2	2270	Where does it say extract info from a line graph?	Framework comment	
2	566	This is just straight multiplication - 4N-2.2 fits best. Rather than understanding different meanings/uses or 2 digit #s.	Framework comment	
2	944	The term "line graph" does not appear in Level 2. If it's really not supposed to be a level 2 skill, then this question is assessing 3S-2.5.	Framework comment	
2	944	Again standard says bar graph - this is line graph	Framework comment	
2	944	I went to lvel 4 - probably have alot of times used 2S-2.3 for getting basic information from a graph, but it does not say line graph like 4S-2.5 does.	Framework comment	
2	944	Same issue as #22. Where does it say extract info from a line graph?	Framework comment	

Level	Item ID	Comment	Code	Details
2	2224	Same note as #27 - line graph not at level 2.	Framework comment	
2	2317	Except it's 3 and 4 digit numbers.	Framework comment	
2	2317	The standard for subtraction is 2 & 3 digits. This is a four digit subtraction.	Framework comment	
3	283	Perhaps benchmark must be moved up.	Framework comment	
3	979	3G-4.11 "measure" used as a label?	Framework comment	
3	979	Standard says measure but measurement is given. We can't assess "measure" on a multiple choice test.	Framework comment	
3	990	Is this really where proportion first appears?	Framework comment	
3	1039	But really 4N-2.1 with very simple numbers.	Framework comment	
3	1467	4P-3.9 should be a level 3 objective.	Framework comment	
3	1467	4P-3.9 should be moved to level 3	Framework comment	
3	1467	Objective 4P-3.9 should be a level 3.	Framework comment	
3	1467	This objective should be in level 3.	Framework comment	
3	2197	Maybe trend should be addressed in level 3.	Framework comment	
3	2363	A very easy level 4	Framework comment	
3	2480	2G-4.3 doesn't cover miles. We need to be more explicit naming appropriate measures to include miles, meters, quarts, liters, etc.	Framework comment	
3	2480	Although there is no mention of miles	Framework comment	

Level	Item ID	Comment	Code	Details
3	2480	Think this is just units of measurement - can't find measurement objective.	Framework comment	
3	374	Objective repeats at level 2 and 3	Framework comment	
3	890	It seems that 3G-4.11 is easier than 2G-4.9 according to enabling skills.	Framework comment	
3	890	Thus was the second problem dealing with squares, which I'm assuming falls under special rectangles.	Framework comment	
3	1851	But this objective should include line graph.	Framework comment	
3	2068	Not as stated in objective but as described in enabling knowledge.	Framework comment	
3	2068	Not sure of objective - is there one that says know horizontal line means division?	Framework comment	
3	2068	Notation - I don't see it	Framework comment	
3	2232	"Friendly numbers" in objective could be inclusive of 20%.	Framework comment	
3	2232	I couldn't find anything that just blanket covered doing percentages without qualifications. Get rid of "using friendly numbers" on 3N-3.12	Framework comment	
3	2404	If this is line graph, too.	Framework comment	
3	2506	Can't find the objective 3S-3.2 plus multiplication.	Framework comment	
4	1015	Standard says graph but here it is a table.	Framework comment	
4	1066	There is no area for level 4.	Framework comment	
4	1443	Could involve +,-,x,integers, ratios. I couldn't find anything that included all operations.	Framework comment	

Level	Item ID	Comment	Code	Details
4	2463	This seems more basic/easier than 3G-1.2	Framework comment	
4	2536	I scatter plots mentioned in the frameworks.	Framework comment	
4	2536	Is "line of best fit" among the general trend stuff or does it hark back into 2?	Framework comment	
4	2536	We didn't consider scatter plots when writing the frameworks - but the "line of best fit" is another measure of central tendency.	Framework comment	
4	2366	The pattern is exponential growth - not stated anywhere I find.	Framework comment	
4	912	I don't see any objective for rounding decimals. 4N-3.1 denotes only practical contexts.	Framework comment	
4	912	This is rounding but not in "practical context".	Framework comment	
4	924	I think 4P-3.6 combines 4P-3.4 and 4P-3.5 Formula - do we consider 2x-8=y a formula?	Framework comment	
4	941	Why don't we have a division strand like 4N-2.2? Why are prime numbers (5N-3.7) not introduced until level 5?	Framework comment	
4	1589	Level 5 - Objective for knowing math terms like consecutive, sum, ??	Framework comment	
4	1823	Missing reciprocal operations strand and applying it to go backwards through equations.	Framework comment	
4	1665	Representing problems in words.	Framework comment	
4	1717	Cannot find pictographs or keys - seems easy for level 4.	Framework comment	
4	1717	The objective doesn't mention pictograph but it might fit.	Framework comment	
4	1717	There is no specific choice for pictograph.	Framework comment	

Level	Item ID	Comment	Code	Details
4	1355	I would guess that the value of coins is a skill lower than level 2.	Framework comment	
4	1355	There is no level 4 money.	Framework comment	
4	1355	This doesn't have to be about money but it does use the money terms.	Framework comment	
4	1355	Value of coins? Does it really fit?	Framework comment	
5	2343	Is there no comparable standard in level 5?	Framework comment	
5	1746	Why isn't level 5 more detailed in higher topics than level 4?	Framework comment	
5	2430	I don't think this is the one, but I can't find one that says write an equation to represent the in/out table.	Framework comment	
2	1719	Why are there 2 relatively easy time clock problems?	Item critique	
2	947	Third clock question	Item critique	
2	1762	Very similar to #4	Item critique	
2	2126	In/Out	Item critique	
2	2212	4th clock question	Item critique	
3	1809	Remove "real life context" from language and state it in question.	Item critique	
3	1809	Terminology distracts from the task at hand.	Item critique	
3	1809	What's a number machine? Is this an issue - enabling skills for 3P-2.2 go to 3P-1.3?	Item critique	
3	1447	I almost didn't notice the key	Item critique	
3	845	Reword to "A family left their house at 11:00 a.m. They..."	Item critique	
4	553	Very easy for level 4 - more about reading than math.	Item critique	

Level	Item ID	Comment	Code	Details
4	930	This should read "By noon, the temperature had risen 20 degrees. At 5pm the temperature had some down 30 degrees. By 8 pm, the temperature had gone down another 8 degrees."	Item critique	
4	985	" a third" instead of "one third" might be confusing.	Item critique	
5	1662	This seems so easy for lvel 5 and #15 on level 4 seemed more difficult. I can't account for why this performed higher.	Item critique	
5	2091	See #3 - very similar?	Item critique	
5	2557	It should say which expression (not equation) because the answers are not equations. Not sure where to put this one! Sort of each of these objectives.	Item critique	
5	2557	Text should say "Which inequality below..."	Item critique	
5	2557	The word "equation" should be changed to expression or inequality!! Equation means equal. Also < before candles in answer choices.	Item critique	
5	1809	Isn't this also on lower level?	Item critique	
5	2247	Not really 5N-1.3 in terms of difficulty.	Item critique	
5	2304	Is this really a level 2 question?	Item critique	
5	2304	Pretty simple for level 5	Item critique	
5	2304	This is level 5!?!	Item critique	
5	1053	Fixed - Pronoun he used along with she for Hilda.	Item critique	
5	1653	Question should say "Which of the following values satify the equation?" because x can be +6 or -6.	Item critique	
5	1676	Why is this simple problem here when students just did on p. 29?	Item critique	

Level	Item ID	Comment	Code	Details
5	1820	It's not clear when the factory shut down, but not necessary for answer.	Item critique	
2	1781	I think the 15 was bold faced instead of the 5 in the answers.	Skip	
2	1781	Wrong answer bolded	Skip	
2	1781	Wrong answer is marked on the sheet.	Skip	
2	1473	3S-2.4 is the same benchmark - what difference is it? Oh - assessed - get it.	Skip	
4	992	Very easy level 4.	Skip	
4	486	This closely resembles a level 3 problem about sisters.	Skip	
4	486	Was the same question in level 3 packet?	Skip	
4	1066	Nice question	Skip	
4	2463	Why can't more of them be this easy to identify?	Skip	
4	2513	This is not what was meant by combination shapes but it is the closest standard.	Skip	
4	2536	Nice question!	Skip	
4	848	Kind of easy - also in level 3.	Skip	
5	2368	I've got an idea! Why don't we estimate before we solve?!	Skip	
2	1719	Could be KC?	Unsure of Cog	
2	2144	Could be ASE?	Unsure of Cog	
2	1037	Possibly ASE? I think App more.	Unsure of Cog	
2	1407	App or ASE? Seems like you have to break down the question.	Unsure of Cog	
3	1449	Could be KC?	Unsure of Cog	
3	1354	The "which" makes me wonder if ASE but seems like App to me.	Unsure of Cog	
3	2197	Maybe App?	Unsure of Cog	

Level	Item ID	Comment	Code	Details
3	2228	Maybe App?	Unsure of Cog	
3	1447	Could be App?	Unsure of Cog	
3	2480	Could be ASE?	Unsure of Cog	
4	2352	ASE because they find the range.	Unsure of Cog	
4	486	ASE because deciding which to choose?	Unsure of Cog	
5	1617	Choose ASE because you have to know how to set up median - put together the numbers, list, then find it. Maybe there should be another objective but I'm not sure.	Unsure of Cog	
5	1617	Could be ASE?	Unsure of Cog	
5	1662	Maybe App? It's hard to decide App/ASE because the thinking on all of these level 5 problems is much more complex - but since they have to come up with the pattern, decide how to get 180, it might be ASE.	Unsure of Cog	
5	932	Questioned if could ASE	Unsure of Cog	
5	1809	Could be App? Maybe ASE because you have to figure out the pattern, then put it into words.	Unsure of Cog	
5	1746	Could be App?	Unsure of Cog	
5	2430	Could be ASE? I think this is a repeat.	Unsure of Cog	
2	1040	Standard - round money to estimate??	Unsure of match	Finding objective
2	1040	They can do this by adding, so I don't know exactly what obj.	Unsure of match	Finding objective
2	566	I know it's not level 4, but I can't find multiply 1 digit by 1 digit effectively.	Unsure of match	Finding objective

Level	Item ID	Comment	Code	Details
2	566	Isn't there a standard at level 2 for one digit by 1 digit multiplication?	Unsure of match	Finding objective
3	990	Is there an objective for scale?	Unsure of match	Finding objective
3	1039	This is a fairly simple problem, but i can't find a level 3 objective that fits.	Unsure of match	Finding objective
3	2197	I'm not sure this is the objective - I couldn't find one that says to make an assumption based on what you have.	Unsure of match	Finding objective
3	2363	Where is "carry out calculations with 2 digit numbers?" Also, I guess this is a 2 step, but it's not a level 4 so I didn't know what to pick?	Unsure of match	Finding objective
3	2232	Hard to find the objective that matches the item.	Unsure of match	Finding objective
3	2232	There is no objective that singley does this.	Unsure of match	Finding objective
3	2282	Is there an objective about reading info from a visual that is not a graph?	Unsure of match	Finding objective
4	1015	There does not seem to be an objective that states compare info from a table, but there is one about a bar graph. Or maybe it's not comparing, but rather just choosing. This one I'm not sure.	Unsure of match	Finding objective
4	2536	I spent way too much time trying to find an objective for this question in both 4 and 5.	Unsure of match	Finding objective
4	2366	Not sure of this objective - it seems more complicated than 4N-2.2 but I can't find one to go with it.	Unsure of match	Finding objective

Level	Item ID	Comment	Code	Details
4	930	It feels like this is a positive/negative number question but I can't find an objective to fit.	Unsure of match	Finding objective
5	2328	This doesn't seem like a level 4 question, but I can't find an objective at 5.	Unsure of match	Finding objective
5	349	These are really hard to pick the correct objective - I can't find exactly what I want.	Unsure of match	Finding objective
2	1009	A bit of a stretch!	Unsure of match	Fit with objectives
2	1040	This is easy if estimation were in 10s but it's not - I think this is harder than level 2.	Unsure of match	Fit with objectives
2	2071	It's not labeled a map, but it is a coordinate grid, so maybe this fits.	Unsure of match	Fit with objectives
2	2071	Not sure 3G-2.3 fits-	Unsure of match	Fit with objectives
3	611	I also like 2G-4.2	Unsure of match	Fit with objectives
3	611	I think this would be 2G-4.2 if time format was used in stead of words.	Unsure of match	Fit with objectives
3	1374	Could be solved with formula or visually.	Unsure of match	Fit with objectives
3	990	4N-3.5 if use ratio and proportion	Unsure of match	Fit with objectives
3	374	3S-5.3 but not ratio	Unsure of match	Fit with objectives
3	890	Not measuring shape but labeling results 3G-4.11	Unsure of match	Fit with objectives
3	2232	Does not necessarily have to use proportion.	Unsure of match	Fit with objectives
4	1671	Very easy level 4. Could be 3S-3.3	Unsure of match	Fit with objectives

Level	Item ID	Comment	Code	Details
4	1420	4G-4.8 is volume, but the question also involves division - better meets multistep problem - 4N-2.1	Unsure of match	Fit with objectives
4	486	Not sure if it is 3 or 4.	Unsure of match	Fit with objectives
4	810	This seems to be my answer when I don't know the answer.	Unsure of match	Fit with objectives
4	1066	I'm not sure if this is a simple multi-step problem or an area problem.	Unsure of match	Fit with objectives
4	930	I don't think this objective is quite right as they are asked to make the calculation not just recognize.	Unsure of match	Fit with objectives
4	1646	It seems a little more - first you have to identify the graph (correctly).	Unsure of match	Fit with objectives
5	936	This also involves reading the odometer!	Unsure of match	Fit with objectives
5	936	This was the closest 5 I could get but really this problem might be more basic?	Unsure of match	Fit with objectives
5	959	This is just independent - the standard states both. Does that mean it really only aligns with part of the standards? Seems 4S-5.4 fits better? 4S-5.3 ratio?	Unsure of match	Fit with objectives
5	2129	Find interest rate - use to calculate or use patterns?	Unsure of match	Fit with objectives
5	1831	5S-4.1 might have addressed two examples on level 4 where we were supposed to choose which graph showed statements.	Unsure of match	Fit with objectives
5	2328	or maybe 4N-3.6 - I don't know!	Unsure of match	Fit with objectives
5	1755	Not sure if I have the right objective.	Unsure of match	Fit with objectives

Level	Item ID	Comment	Code	Details
5	349	This is a difficult question, but I don't know if it is assessing fractions or division or estimating.	Unsure of match	Fit with objectives
2	1009	Not sure if this is skip counting or something else...	Unsure of match	Not sure
3	1449	Really note sure where this question fits.	Unsure of match	Not sure
3	845	I'm not really sure.	Unsure of match	Not sure
4	2323	Seems to be more to it. Not sure.	Unsure of match	Not sure
4	1589	Beats the heck out of me! My lack of knowledge is source of challenge!	Unsure of match	Not sure
4	1589	I'm not sure.	Unsure of match	Not sure
4	1749	I don't really know.	Unsure of match	Not sure
5	2557	Not at all sure of this one.	Unsure of match	Not sure
2	2209	If calculator is provided 2P-3.1	Unsure of match	Skip
3	1449	I didn't know which to choose becaue you have to do both. I think each is equal.	Unsure of match	Skip
4	1420	I think you need both standards.	Unsure of match	Skip
4	1846	I chose this one because I think it means how does a change in the data affect the mean.	Unsure of match	Skip
5	1664	I think it's this because you have to evaluate each formula compared to the info given.	Unsure of match	Skip

APPENDIX H

FRAMEWORK MODIFICATIONS BY LEVEL

Level 2
Add objectives about:

- Decimals
- Multi-step operations with simple numbers
- Line graphs
- Operations - addition of basic sums to 10, to 20, two digit with carrying, subtraction with and without borrowing, subtraction when zeros are involved, etc.
- Straight one digit by one digit multiplication (like 4N-2.2)

Level 3
Add objectives about:

- Scale
- Using a number line to represent values (switch 4P-3.9 and 3P-3.7)
- Trend
- Calculations with 2 digit numbers just clearly stated
- Line graphs
- Straight percentage calculations (not just with friendly numbers)
- Pictographs

Level 4
Add objectives about:

- Comparing information from a table not just from a graph
- All operations
- Line of best fit and how it fit with trends
- Exponential growth patterns
- Rounding decimals (not necessarily in a practical context)
- Performing division operations reliably, accurately, and efficiently.
- Prime number
- Specific math terms – consecutive, sum, etc.
- Pictographs
- Finding percentages
- 4P-3.6 may combine 4P-3.5 and 4P-3.4

Level 5

- Add objectives about:
- Multi-step operations
- Equations for in/out tables

APPENDIX I

TOPICAL VIEW OF THE ITEMS WITHIN EACH LEVEL

Path	Level 2

Topic	Item	Total	Average	
Addition	10	1		
	31	1		
	32	2		
Addition Total		4	0.67	
Addition/ Subtraction	3	4		
	5	1		
	19	1		
	23	2		
	31	1		
	37	4		
	38	2		
Addition/Sub Total		15	2.50	
Data - Collect, Org, Rep	8	1		
	22	2		
	29	1		
Data - Collect, Organize, and Represent Total		4	0.67	
Data - Description, Statistics, Trends	10	1		
Data - Description, Statistics, Trends Total		1	0.17	Added
Data - Make and Evaluate Statements by Applying Knowledge of Data	22	1		
	29	3		
Data - Make and Evaluate Statements by Applying Knowledge of Data Total		4	0.67	
Decimal/ Percentage/ Fraction - Equivalent	13	1		
	39	1		
Decimal/ Percentage/ Fraction - Equivalent Total		2	0.33	Added
Decimals	9	2		
	28	4		
Decimals Total		6	1.00	Added

Missing topics

Number Sense	Calculator usage
Patterns	Missing variables
	Number line
	Place value/ineq.
Geometry	Area/peri./volume
	Shapes-properties

Topic	Item	Total	Average
Division	3	1	
	6	1	
	10	5	
	13	1	
Division Total		8	1.33
Fractions	1	6	
	6	5	
	39	1	
Fractions Total		12	2.00
Map skills/ Coordinates	17	6	
	34	6	
	35	6	
Map skills/ Coordinates Total		18	3.00
Measurement	8	2	
	40	1	
Measurement Total		3	0.50
Measurement equivalency	12	6	
Measurement equivalency Total		6	1.00
Money	3	1	
	9	2	
	14	6	
	19	6	
	38	1	
Money Total		16	2.67
Multiplication	9	3	
	15	3	
	26	5	
	37	1	
Multiplication Total		12	2.00
Multiplication/Division	3	1	
	26	1	
Multiplication/Division Total		2	0.33
Operations - general	10	1	
	15	3	
	31	5	
Operations - general Total		9	1.50
Pattern - identification	18	4	
	23	5	
	30	2	
	37	3	
Pattern - identification Total		14	2.33

Topic	Item	Total	Average
Patterns - Represent Relationships with Tables, Graphs, Rules, Equations	9	2	
	18	3	
	30	4	
Patterns - Represent Relationships with Tables, Graphs, Rules, Equations Total		9	1.50
Percentages	13	5	
	39	4	
Percentages Total		9	1.50
Place value	8	2	
	19	1	
	21	1	
	28	2	
	33	6	
Place value Total		12	2.00
Probability	30	2	
Probability Total		2	0.33
Read and interpret	4	6	
	5	3	
	8	1	
	10	1	
	16	6	
	21	6	
	22	3	
	24	6	
	27	6	
	29	2	
	32	5	
	36	6	
Read and interpret Total		51	8.50
Rounding/ Estimation	9	2	
Rounding/ Estimation Total		2	0.33
Subtraction	5	3	
	25	7	
	31	1	
	32	1	
	38	4	
Subtraction Total		16	2.67
Symmetry	6	2	
Symmetry Total		2	0.33
Temperature	40	5	
Temperature Total		5	0.83

Topic	Item	Total	Average
Time	2	6	
	7	6	
	11	6	
	18	1	
	20	6	
Time Total		25	4.17
Grand Total		269	44.83

Path	Level 3

Topic	Item	Total	Average
Addition/ Subtraction	11	5	
	14	1	
	37	1	
	39	1	
Addition/ Subtraction Total		8	1.33
Area/perimeter/volume	5	7	
	12	6	
	30	7	
Area/peri./volume Total		20	3.33
Calculator usage	2	1	
Calculator usage Total		1	0.17
Data - Collect, Organize, and Represent	23	1	
Data - Collect, Organize, and Represent Total		1	0.17
Data - Description, Statistics, Trends	38	1	
Data - Description, Statistics, Trends Total		1	0.17
Data - Make and Evaluate Statements by Applying Knowledge of Data	19	2	
	33	5	
	38	1	
Data - Make and Evaluate Statements by Applying Knowledge of Data Total		8	1.33
Decimal/ Percentage/ Fraction - Equivalent	9	1	
	17	4	
	34	1	
Decimal/ Percentage/ Fraction - Equivalent Total		6	1.00
Decimals	17	1	
Decimals Total		1	0.17
Division	3	4	
	28	1	
Division Total		5	0.83

Missing Topics

Number Sense	Exponents/roots
	Fractions – manip.
Geometry	Symmetry
	Triangles/Angles

Topic	Item	Total	Average	
Fractions	9	5		
	15	6		
	22	2		
	34	1		
Fractions Total		14	2.33	
Map skills/ Coordinates	8	6		
	13	2		
	24	1		
Map skills/ Coordinates Total		9	1.50	
Measurement	24	4		
Measurement Total		4	0.67	
Measurement equivalency	24	1		
Measurement equivalency Total		1	0.17	
Missing Variables	10	6		
	29	2		
	30	1		
Missing Variables Total		9	1.50	
Money	2	4		
	6	1		
	11	6		
	37	7		
Money Total		18	3.00	
Multiplication	13	1		
	14	1		
	22	1		
	23	1		
Multiplication Total		4	0.67	Added
Multiplication/Division	3	1		
	17	1		
Multiplication/Division Total		2	0.33	
Number Line	16	7		
	25	2		
Number Line Total		9	1.50	
Operations - general	2	2		
	6	5		
	7	1		
	14	1		
	22	4		
	27	6		
	29	1		
	34	3		
	40	1		
Operations - general Total		24	4.00	

Topic	Item	Total	Average	
Pattern - identification	14	2		
	21	3		
	40	2		
Pattern - identification Total		7	1.17	
Patterns - Represent Relationships with Tables, Graphs, Rules, Equations	14	1		
	19	1		
	21	3		
	40	3		
Patterns - Represent Relationships with Tables, Graphs, Rules, Equations Total		8	1.33	
Percentages	31	2		
	36	7		
Percentages Total		9	1.50	
Place value	1	6		
	18	2		
	25	1		
	32	6		
Place value Total		15	2.50	
Place value - inequality	18	5		
Place value - inequality Total		5	0.83	
Probability	26	6		
	31	4		
Probability Total		10	1.67	
Probability - ratio	26	1		
	31	1		
Probability - ratio Total		2	0.33	
Rate of change	29	3		
Rate of change Total		3	0.50	
Ratio	13	3		
Ratio Total		3	0.50	Added
Read and interpret	7	9		
	19	3		
	23	4		
	33	1		
	35	6		
	38	5		
	39	5		
Read and interpret Total		33	5.50	
Rounding/ Estimation	3	1		
Rounding/ Estimation Total		1	0.17	

Topic	Item	Total	Average
Shapes - properties	5	2	
	12	1	
	20	6	
Shapes - properties Total		9	1.50
Subtraction	19	1	
Subtraction Total		1	0.17
Temperature	25	4	
Temperature Total		4	0.67
Time	4	6	
	28	6	
	29	1	
Time Total		13	2.17
Grand Total		268	44.67

Added

Path	Level 4

Topic	Item	Total	Average
Addition/ Subtraction	2	1	
	16	1	
	29	1	
Addition/ Subtraction Total		3	0.50
Area/perimeter/volume	4	6	
	6	1	
	10	5	
	14	6	
	22	7	
	38	5	
Area/peri./volume Total		30	5.00
Data - Collect, Organize, and Represent	36	1	
Data - Collect, Organize, and Represent Total		1	0.17
Data - Description, Statistics, Trends	7	6	
	13	3	
	18	7	
	23	2	
	30	1	
	34	8	
Data - Description, Statistics, Trends Total		27	4.50
Data - Make and Evaluate Statements by Applying Knowledge of Data	17	4	
	40	4	
Data - Make and Evaluate Statements by Applying Knowledge of Data Total		8	1.33
Decimals	27	5	
Decimals Total		5	0.83
Exponents/ roots	3	6	
	9	4	
	20	6	
	24	1	
Exponents/ roots Total		17	2.83

Missing Topics

Number Sense	Calculator usage
	Dec./Perc/Frac – Eq
	Division
	Fractions
Patterns	Place value - ineq
Statistics	Probability
	Probability - ratio
Geometry	Shapes - properties
	Measurement
	Circles

Topic	Item	Total	Average	
Fractions - manipulating	31	6		
	36	1		
Fractions - manipulating Total		7	1.17	
Integers	19	2		
	29	5		
Integers Total		7	1.17	
Map skills/ Coordinates	37	5		
Map skills/ Coordinates Total		5	0.83	
Measurement equivalency	10	1		
Measurement equivalency Total		1	0.17	
Missing Variables	4	1		
	9	1		
	28	8		
	33	1		
Missing Variables Total		11	1.83	
Money	39	6		
Money Total		6	1.00	Added
Multiplication	24	2		
Multiplication Total		2	0.33	
Multiplication/Division	6	3		
	12	1		
	25	1		
	30	4		
Multiplication/Division Total		9	1.50	
Number Line	26	2		
Number Line Total		2	0.33	
Operations - general	2	5		
	5	6		
	6	1		
	9	2		
	10	2		
	11	8		
	12	5		
	14	1		
	16	4		
	19	5		
	22	1		
	25	1		
	29	1		
	33	5		
	34	1		
	35	6		
Operations – gen Total		54	9.00	

Topic	Item	Total	Average
Pattern - identification	8	4	
	13	1	
	15	5	
	16	2	
	24	3	
	29	1	
	32	6	
	33	1	
Pattern - identification Total		23	3.83
Patterns - Represent Relationships with Tables, Graphs, Rules, Equations	8	1	
	14	1	
	15	2	
	17	1	
	18	1	
	23	2	
Patterns - Represent Relationships with Tables, Graphs, Rules, Equations Total		8	1.33
Percentages	25	1	
	31	1	
Percentages Total		2	0.33
Place value	1	6	
Place value Total		6	1.00
Rate of change	8	2	
	23	1	
	25	5	
	30	2	
Rate of change Total		10	1.67
Ratio	6	4	
	10	1	
	31	1	
Ratio Total		6	1.00
Read and interpret	13	5	
	17	1	
	23	2	
	36	6	
	37	1	
	40	1	
Read and interpret Total		16	2.67
Rounding/ Estimation	18	1	
Rounding/ Estimation Total		1	0.17
Symmetry	23	1	
Symmetry Total		1	0.17

Topic	Item	Total	Average
Temperature	26	4	
	29	1	
	38	1	
Temperature Total		6	1.00
Triangles/ Angles	21	6	
Triangles/ Angles Total		6	1.00
Grand Total		280	46.67

Path	Level 5

Topic	Item	Total	Average
Addition/ Subtraction	2	2	
	7	1	
	37	1	
Addition/ Sub.Total		4	0.67
Area/perimeter/volume	5	5	
	22	4	
	33	1	
	39	1	
Area/peri./volume Total		11	1.83
Calculator usage	11	2	
	20	2	
	23	1	
	29	1	
Calculator usage Total		6	1.00
Data - Collect, Organize, and Represent	13	1	
	26	1	
Data - Collect, Organize, and Represent Total		2	0.33
Data - Description, Statistics, Trends	4	6	
	23	6	
	29	6	
	32	6	
Data - Description, Statistics, Trends Total		24	4.00
Data - Make and Evaluate Statements by Applying Knowledge of Data	13	3	
	17	3	
	25	1	
	26	1	
	28	2	
	35	3	
	36	2	
Data - Make and Evaluate Statements by Applying Knowledge of Data Total		15	2.50

Missing Topics
Number Sense
Division
Fractions – Manip.

Added

Topic	Item	Total	Average	
Decimal/ Percentage/ Fraction - Equivalent	27	7		
Decimal/ Percentage/ Fraction - Equivalent Total		7	1.17	
Decimals	9	3		
	18	2		
Decimals Total		5	0.83	
Exponents/ roots	20	1		
	21	6		
	31	1		
Exponents/ roots Total		8	1.33	
Fractions	26	1		
Fractions Total		1	0.17	
Integers	1	3		
Integers Total		3	0.50	
Map skills/ Coordinates	8	6		
Map skills/ Coordinates Total		6	1.00	
Measurement	5	1		
	22	1		
	33	1		
Measurement Total		3	0.50	
Measurement equivalency	18	3		
Measurement equivalency Total		3	0.50	Added
Missing Variables	12	1		
	14	2		
	15	1		
	20	5		
	24	1		
	26	1		
	31	5		
	39	1		
	40	1		
Missing Variables Total		18	3.00	
Multiplication	2	2		
Multiplication Total		2	0.33	Added
Multiplication/Division	20	1		
Multiplication/Division Total		1	0.17	Added

Topic	Item	Total	Average	
Operations - general	2	2		
	6	1		
	9	3		
	14	4		
	15	1		
	23	1		
	24	4		
	40	1		
Operations - general Total		17	2.83	
Pattern - identification	6	5		
	7	4		
	15	1		
	16	5		
	28	1		
	37	5		
	40	1		
Pattern - identification Total		22	3.67	
Patterns - Represent Relationships with Tables, Graphs, Rules, Equations	7	1		
	13	1		
	16	2		
	17	2		
	24	1		
	28	2		
	35	1		
	40	3		
Patterns - Represent Relationships with Tables, Graphs, Rules, Equations Total		13	2.17	
Percentages	7	1		
	11	7		
	25	2		
Percentages Total		10	1.67	
Place value	1	2		
Place value Total		2	0.33	
Place value - inequality	15	4		
Place value - inequality Total		4	0.67	Added
Probability	3	7		
	10	6		
	30	6		
	38	5		
Probability Total		24	4.00	

Topic	Item	Total	Average	
Probability - ratio	3	1		
Probability - ratio Total		1	0.17	Added
Ratio	12	6		
	18	2		
	26	1		
	38	1		
Ratio Total		10	1.67	
Read and interpret	13	3		
	17	1		
	19	1		
	20	1		
	25	3		
	28	2		
	35	2		
	36	4		
	37	1		
Read and interpret Total		18	3.00	
Rounding/ Estimation	26	1		
Rounding/ Estimation Total		1	0.17	Added
Shapes - properties	33	5		
	34	1		
Shapes - properties Total		6	1.00	
Subtraction	1	1		
Subtraction Total		1	0.17	Added
Temperature	19	4		
	22	1		
Temperature Total		5	0.83	
Triangles/ Angles	34	6		
	39	4		
Triangles/ Angles Total		10	1.67	
Grand Total		263	43.83	

BIBLIOGRAPHY

Ananda, S. (2003a). Achieving alignment. *Leadership, 33*(1), 18-22.

Ananda, S. (2003b). *Rethinking issues of alignment under "No Child Left Behind."* San Francisco, CA: WestEd.

Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory into Practice, 41*(4).

Belzer, A., Drennon, C., & Smith, C. (2001). Building professional development systems in adult basic education: Lessons from the field. Retrieved Dec. 10, 2006, from http://www.ncsall.net/?id=559

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Blank, R. K. (2004, April). *Findings on alignment of instruction using enacted curriculum data: Results from urban schools.* Paper presented at the Annual meeting of American Educational Research Association, San Diego, CA.

Blank, R. K., Porter, A. C., & Smithson, J. L. (2001). *New tools for analyzing teaching, curriculum and standards in Mathematics and Science.* Washington, D.C.: Council of Chief State Schools Officers.

Bloom, B. S. (1956). *Taxonomy of educational objectives, Handbook I: The cognitive domain.* New York: David McKay Co Inc.

Borko, H. (1997). New forms of classroom assessment: Implications for staff development. *Theory into Practice, 36*(4), 231-238.

Borko, H., Davinroy, K. H., Bliem, C. L., & Cumbo, K. (2000). Exploring and supporting teacher change: Two third-grade teachers' experiences in a mathematics and literacy staff development project. *The Elementary School Journal, 100*(4), 273-306.

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13*(3), 259-278.

Buckendahl, C. W., Plake, B. S., Impara, J. C., & Irwin, P. M. (2000, April). *Alignment of standardized achievement tests to state content standards: A comparison of publishers' and teachers' perspectives.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Cavanagh, S. (2004). Bush takes on critics of No Child Left Behind Act. *Education Week, 23*(37), 28.

CCSSO. (2002). Models for alignment analysis and assistance to states. Retrieved August 28, 2005, from www.ccsso.org/content/pdfs/AlignmentModels.pdf

CCSSO. (2005). Models. Retrieved August 28, 2005, from http://www.ecs.org/html/offsite.asp?document=http%3A%2F%2Fwww%2Eccsso%2Eorg%2Fprojects%2FAlignment%5FAnalysis%2FModels%2F

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19-27.

Cohen, D. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*(3), 3111-3344.

Cohen, D. K. (1991). Revolution in One Classroom (or, Then Again, Was It?). *American Educator: The Professional Journal of the American Federation of Teachers, 15*(2), 16-23,44-48.

Comings, J., & Soricone, L. (2005). Massachusetts: A case study of improvement and growth in adult education services. Retrieved Dec. 10, 2006, from http://www.ncsall.net/fileadmin/resources/ann_rev/rall_v5_ch4.pdf

Comings, J., Sum, A., & Uvin, J. (2000). *New skills for a new economy: Adult education's key role in sustaining economic growth and expanding opportunity*. Boston: Massachusetts Institute for a New Commonwealth (MassInc).

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Crocker, L. M. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement, Issues and Practice, 22*(3), 5-11.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, D.C.: American Council on Education.

Darling-Hammond, L. (2003). Standards and assessments: Where we are and what we need. 2003, from http://www.tcrecord.org ID Number 11109

Falk, B., & Ort, S. (1997). *Sitting down to score: Teacher learning through assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Franke, M., Carpenter, T., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: A follow-up study of professional development in Mathematics. *American Educational Research Journal, 38*(3), 653-689.

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*(4), 915-945.

Guskey, T. (2005). Mapping the road to proficiency. *Educational Leadership, 63*(3), 32-38.

Herman, J. (2002). *Instructional effects in Elementary schools* (No. CSE Technical Report 577). Los Angeles, CA: CRESST/University of California, Los Angeles.

Herman, J., Webb, N., & Zuniga, S. (2005). *Measurement issues in the alignment of standards and assessments: A case study* (No. CSE Report 653). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

James, K. W. (2004). A closer look at the underlying issues of the No Child Left Behind Act. *Board & Administrator; for Superintendents Only, 18*(2), 1-4.

Johnson, R. B. (2006). Chapter 14 Mixed Research: Mixed Method and Mixed Model Research. Retrieved March 17, 2006, from http://www.southalabama.edu/coe/bset/johnson/dr_johnson/lectures/lec14.htm

Kauffman, D., Johnson, S. M., Kardos, S. M., Liu, E., & Peske, H. G. (2002). "Lost at Sea": New teachers' experiences with curriculum and assessment. *Teachers College Record, 104*(2), 273-200.

Lane, S., Stone, C., Parke, C. S., Hansen, M. A., & Cerrillo, T. L. (2000, April). *Consequential evidence for MSPAP from the teachers, principal and student perspective.* Paper presented at the Annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Le Marca, P. M. (2001). *Alignment of standards and assessments as an accountability criterion. Eric Digest* (No. ED458288): ERIC Development Team.

Le Marca, P. M., Redfield, D., Winter, P. C., & Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment. Series on standards and assessments*. Washington, D.C.: Council of Chief State Schools Officers.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership, 63*(3), 19-24.

Leffler, J. C., Carr, M., Griffin, L., & Gates, C. (2005). *Alignment of Montana state standards with state assessments*. Portland, OR: Northwest Regional Educational Laboratory.

Leffler, J. C., Potter, J., Novick, R., Carr, M., Gates, C., Leong, M., et al. (2003). *Alignment of Idaho state standards with Idaho Standards Achievement Test (ISAT)*. Portland, OR: Northwest Regional Educational Laboratory.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Love, N. (2000). *Using data-getting results: Collaborative inquiry for school-based mathematics and science reform*. Cambridge, MA: TERC.

Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. (Massachusetts Comprehensive Assessment System). *English Journal, 91*(1), 79-87.

Martone, A. (2005, April). *"Making the private work of teaching public": The role of a full-time math teacher leader for a school labeled underperforming.* Paper presented at the American Educational Research Association, Montreal, Canada.

Martone, A., Goodridge, B., Moses, M., & Titzel, J. (2004). *Refinements to ABE mathematics standards for assessment purposes* (No. 548). Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Martone, A., Sireci, S. G., & Delton, J. (2007). *Methods for evaluating the alignment between state curriculum frameworks and state assessments: A literature review* (No. 603). Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

McGehee, J. J., & Griffith, L. K. (2001). Large-scale assessments combined with curriculum alignment: Agents of change. *Theory into Practice, 40*(2).

McTighe, J., & O'Connor, K. (2005). Seven practices for effective learning. *Educational Leadership, 63*(3), 10-17.

Mullinix, B. (1994). *Exploring what counts: Mathematics instruction in adult basic education*. Boston: World Education.

Petit, M. (2002, October). *Improving the No Child Left Behind Act.* Paper presented at the Reidy Interactive Lecture Series, Nashua, NH.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5*, 285-301.

Popham, W. J. (2004). Curriculum, instruction, and assessment: Amiable allies or phony friends. Retrieved November 2, 2005, from http://www.tcrecord.org/PrintContent.asp?ContentID=11522

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Porter, A. C., & Smithson, J. L. (2001). *Defining, Developing, and Using Curriculum Indicators. CPRE Research Report Series* (No. RR-048). Philadelphia, PA: Consortium for Policy Research in Education.

Porter, A. C., & Smithson, J. L. (2002, April). *Alignment of assessments, standards and instruction using curriculum indicator data.* Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, LA.

Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin Alternate Assessment. *The Journal of Special Education, 38*(4), 218-231.

Rossman, G., & Rallis, S. (2003). *Learning in the field: An introduction to qualitative research*. Thousand Oaks, CA: Sage.

Rothman, R. (2003). *Imperfect matches: The alignment of standards and tests*: National Research Council.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing. CSE Technical Report* (No. CSE-TR-566). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Rothstein, R. (2002, May 1). States teeter when balancing standards with tests. *New York Times.*

Scherer, M. (2005). Reclaiming testing. *Educational Leadership, 63*(3), 9.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 2*, 420-428.

Sireci, S. G. (1998a). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Sireci, S. G. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299-321.

Sireci, S. G., Baldwin, P., Keller, L., Valle, M., Goodridge, B., Leonelli, E., et al. (2004). *Specifications for the ACLS Mathematics and Numeracy Proficiency Tests* (No. 513). Amherst, MA: Center for Educational Assessment, School of Education, University of Massachusetts Amherst.

Sireci, S. G., Baldwin, P., Martone, A., Laguilles, A. Z., Hambleton, R. K., & Han, K. T. (2006). *Massachusetts Adult Proficiency Tests: Technical manual* (No. 600). Amherst, MA: Center for Educational Assessment, School of Education, University of Massachusetts Amherst.

Sireci, S. G., Robin, F., Meara, K., Rogers, H. J., & Swaninathan, H. (2000). An external evaluation of the 1996 Grade 8 NAEP Science Framework. In N. Raju, J. W. Pelligrino, M. W. Bertenthal, K. J. Mitchell & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 74-100). Washington, D.C.: National Academies Press.

Slattery, J. B., personal communication, December 15, 2006

Smith, C., & Hofer, J. (2003). The characteristics and concerns of adult basic education teachers. Retrieved Dec 10, 2006, from http://www.ncsall.net/fileadmin/resources/research/brief26.pdf

Smith, M., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233-267). Bristol, PA: Taylor & Francis.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 4*(10), 7-11.

Sparks, G. M. (1988). Teachers' attitudes toward change and subsequent improvements in classroom teaching. *Journal of Educational Psychology, 80*(1), 111-117.

Stecher, B., Barron, S. L., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classrooms* (No. CSE Tech. Rep. 525). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B., & Borko, H. (2002). *Combining surveys and case studies to examine standards-based educational reform* (No. CSE. Tech. Rep. 565). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Stipek, D., Givvin, K. B., Salmon, J. M., & Macgyvers, V. L. (1998). Can a teacher intervention improve classroom practices and student motivation in mathematics? *Journal of Experimental Education, 66*(4), 319-337.

Swafford, J. O., Jones, G. A., Thornton, C. A., Stump, S. L., & Miller, D. R. (1999). The impact on instructional practice of a teacher change model. *Journal of Research and Development in Education, 32*(2), 69-82.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago: The University of Chicago Press.

Webb, N., Herman, J., & Webb, N. L. (2006). *Alignment of mathematics state-level standards and assessments: The role of reviewer agreement* (No. CSE Report 685). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Webb, N. L. (1997). *Research monograph no. 6: Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, D.C.: Council of Chief State Schools Officers.

Webb, N. L. (1999). *Research monograph no. 18: Alignment of Science and Mathematics standards and assessments in four states*. Washington, D.C.: Council of Chief State Schools Officers.

Webb, N. L. (2002, April 1-5). *An analysis of the alignment between Mathematics standards and assessments for three states*. Paper presented at the American Educational Research Association, New Orleans, LA.

Webb, N. L., Alt, M., Ely, R., Cormier, M., & Vesperman, B. (2005). *The WEB alignment tool: Development, refinement, and dissemination*. Washington, D.C.: Council of Chief State School Officers.

Webb, N. L., Alt, M., Ely, R., & Vesperman, B. (2005). Web alignment tool (WAT): Training manual draft 1.1. Retrieved March 17, 2006, from http://www.wcer.wisc.edu/WAT/Training%20Manual%202.1%20Draft%20091205.doc

Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. *Journal of Negro Education, 62*, 288-310.

Wise, L. L., & Alt, M. (2005). *Assessing vertical alignment.* Washington, D.C.: Council of Chief State School Officers.

Wolf, D. P., & White, A. M. (2000). Charting the course of student growth. *Educational Leadership, 57*(5).