9-1-2010

Using Item Mapping to Evaluate Alignment between Curriculum and Assessment

Leah Tepelunde Kaira
University of Massachusetts - Amherst, leahkaira@gmail.com

USING ITEM MAPPING TO EVALUATE ALIGNMENT BETWEEN CURRICULUM AND ASSESSMENT

A Dissertation Presented

by

Leah T. Kaira

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

DOCTOR OF EDUCATION

September 2010

Education

USING ITEM MAPPING TO EVALUATE ALIGNMENT BETWEEN CURRICULUM AND ASSESSMENT

A Dissertation Presented

by

LEAH T. KAIRA

Approved as to style and content by:

_____
Stephen Sireci, Chair

_____
Aline Sayer, Member

_____
Ronald Hambleton, Member

_____
Craig S. Wells, Member

_____
April Zenisky, Member

_____
Christine B. McCormick, Dean
School of Education

DEDICATION

To Peter, Alice, Samuel, Linly, and Loraine

ACKNOWLEDGMENTS

Writing this dissertation was made easier because of the support of many people. My deep gratitude goes to my committee members whose guidance helped me refine this dissertation. I am grateful to my advisor and chair, Professor Stephen Sireci who was always there for me. His guidance, careful reading of the dissertation, and dedication were instrumental to successful completion of my work. I truly cannot thank him enough.

This dissertation required knowledge of item response theory and some statistical skills. I therefore extend special thanks to Professor Ronald Hambleton who taught me item response theory with a lot of hands on experiences. I extend my gratitude to Professors Lisa Keller, Craig Wells and Aline Sayer who instilled in me the statistical skills I needed for this dissertation. I really appreciate their great work.

My gratitude also goes to Professors Lisa Keller, April Zenisky, Craig Wells, and Stephen Sireci who were always there for me when I needed emotional support. Their word of encouragement brought a lot of inner comfort and made me carry on with my work with renewed strength. I also thank brothers and sisters of the New Apostolic Church (Springfield congregation) for their unwavering love toward me and my family. They opened their homes to my family and supported us spiritually. I always found comfort in their open hearts. I thank them sincerely. This dissertation was possible because of the support of many close friends. I am especially grateful to Lynn Shelley Sireci, Evans Tchongwe, and Sarah Kahando for all their support and love.

Very special thanks go to my family. My loving husband Samuel has always been there for me. He has given me all anyone could ask for on this journey. I thank him for all that he had to go through for the sake of our family. I would not have achieved what I have without his love, patience, care and understanding. Our loving daughters Linly and

Loraine had to deal with many issues throughout my studies. I owe my thanks to them also for their support.

Lastly, I thank my parents Peter and Alice Tepelunde for the step they took to send me to school. They sacrificed a lot to put resources together for the sake of my education. I am grateful for the great lessons they have taught me through the years and the values I have acquired from them. I have no doubt they are so proud of me and my success.

ABSTRACT

USING ITEM MAPPING TO EVALUATE ALIGNMENT BETWEEN CURRICULUM AND ASSESSMENT

SEPTEMBER 2010

LEAH T. KAIRA, B.Ed., UNIVERSITY OF MALAWI

M.Ed., UNIVERSITY OF MASSACHUSETTS AMHERST

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Stephen G. Sireci

There is growing interest in alignment between state's standards and test content partly due to accountability requirements of the No Child Left Behind (NCLB) Act of 2001. Among other problems, current alignment methods almost entirely rely on subjective judgment to assess curriculum-assessment alignment. In addition none of the current alignment models accounts for student actual performance on the assessment and there are no consistent criteria for assessing alignment across the various models. Due to these problems, alignment results employing different models cannot be compared. This study applied item mapping to student response data for the Massachusetts Adult Proficiency Test (MAPT) for Math and Reading to assess alignment. Item response theory (IRT) was used to locate items on a proficiency scale and then two criterion response probability (RP) values were applied to the items to map each item to a proficiency category. Item mapping results were compared to item writers' classification of the items. Chi-square tests, correlations, and logistic regression were used to assess the degree of agreement between the two sets of data. Seven teachers were convened for a one day meeting to review items that do not map to intended grade level to explain the

misalignment. Results show that in general, there was higher agreement between SMEs classification and item mapping results at RP50 than RP67. Higher agreement was also observed for items assessing lower level cognitive abilities. Item difficulty, cognitive demand, clarity of the item, level of vocabulary of item compared to reading level of examinees and mathematical concept being assessed were some of the suggested reasons for misalignment.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

One component of most educational systems is student assessment. Among other reasons, assessments are put in place to judge and monitor the quality of student learning and for accountability purposes. Accurate evaluation of student learning can be achieved only if there is agreement among the curriculum, what the students learn, and what appears on the assessment. Similarly, assessment results are useful for accountability purposes if the assessment mirrors the curriculum. Therefore there is a need to ensure that there is agreement between the curriculum and the assessment for valid inferences to be drawn from assessment results.

One strategy used to evaluate the match between the curriculum and the assessment is carrying out alignment studies. Bhola, Impara, and Buckendahl (2003) define alignment as "the degree of agreement between a state's content standards for a specific subject and the assessment(s) used to measure student achievement of these standards" (p. 21). It is noted from this definition that the goal of alignment is to establish the degree of match between test content and subject area content as specified in the standards. It is important to emphasize the words 'degree of agreement' because as La Marca, Redfield, Winter, Bailey and Despriet (2000) noted, "It is improbable that a single assessment instrument will provide the breadth of coverage necessary for an aligned system" (p. 18). Porter (2002) also explained that "… tests are a sample of items from the domain, whereas the standards represent the domain. Tests are therefore not expected to cover every content standard but instead are expected to cover a representative sample of the content standards in order to make valid generalizations to the content domain

defined by the standards" (p.1). An alignment study would therefore show the extent to which the content on the standards has been covered by the assessment.

There is growing interest in alignment between state's standards and test content partly for accountability purposes. Under the No Child Left Behind Act of 2001 (NCLB), states are required to have assessments that are aligned to the standards for each subject and grade level. NCLB also requires states using norm-referenced testing to carry out an alignment study to identify state standards omitted in the assessment (Webb, Cormier, & Vesperman, 2005). Even more important is the fact that rewards and sanctions are imposed on states based on assessment results. The high-stakes nature of the consequences associated with performance on tests has led educators to focus their attention on improving student learning and eventually improved performance on tests. Such high stakes associated with test scores demand that sufficient evidence be provided to support particular use of test scores. Research on curriculum- assessment alignment is therefore important for states to fulfill requirements of NCLB.

It can be inferred from the discussion above that alignment is closely related to inferences drawn based on test scores. According to the *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*) (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), drawing correct inferences from test scores is an issue of validity. The *Standards* define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests" (p. 9). Validation is therefore a process of collecting evidence to support the type of inferences that are drawn from test scores. Results of an alignment study can thus be used as validity evidence to support the interpretation of test scores. As Ananda (2003a) stated,

alignment could provide three sources of validity evidence: content, construct, and consequential. Alignment could be a source of content validity evidence because it seeks to establish the degree to which the test reflects the curriculum. In validity studies, content congruence between an assessment and the curriculum is evaluated in terms of domain definition, domain representativeness, domain relevance, and appropriateness of test construction procedures (Sireci, 1998). Domain definition refers to specification of the content and processes to be measured (Thorndike, 1997). This specification involves operationally defining the content to be assessed and making explicit the importance or meaningfulness of the construct represented by the content. Subject matter experts (SMEs) can be used to evaluate the operational definition of a test, which is usually in the form of test specifications.

Representativeness refers to the degree to which items on the test sample the specified content domain (Crocker & Algina, 1986). According to Haynes, Richard and Kubany (1995), a test is considered representative to the degree that the entire domain of the targeted construct can be reproduced. Domain representation also assesses the proportion of test items allocated to each content area or standard and each cognitive process. Data regarding domain representation are typically gathered by asking SMEs to review the test specifications and test items and have them match each item to the content and skills dimensions that make up the test specifications. This analysis establishes the degree to which the emphasis in the assessment corresponds to the emphasis stipulated in the test specifications for each content and skill area. When domain representation is established, it is inferred that the examinee would perform with the same proficiency on a test containing items like those on the validated test.

Relevance in content validity studies refers to the appropriateness of items in a test for the targeted construct and function of the assessment (Haynes et al., 1995). In assessing relevance, test items are judged to establish the extent to which they measure the construct that they intend to measure. In a traditional content validity study SMEs rate the degree to which an item is relevant to its objective or to the job using an ordinal relevance rating scale (e.g., 0=not at all relevant, 6=very relevant). Appropriateness of test development procedures refers to all processes used when constructing a test to ensure that test content fully represents the construct intended to be measured and does not measure irrelevant material (Martone, Sireci & Delton, 2006). Among the four aspects of content validity evidence outlined above, alignment studies could provide evidence about domain representation and relevance.

Second, construct validity evidence involves establishing the extent to which an assessment accurately measures the concepts it is supposed to measure. Alignment studies provide construct validity evidence by showing the progression in complexity in the assessment of a particular concept across grade levels. Finally, alignment is related to consequential validity in that it also seeks to evaluate the social consequences of an assessment such as improved student learning (Ananda, 2003a), and the degree to which the intended curriculum is implemented.

1.2 Issues with Current Methods of Alignment

For over a decade, research in alignment has not only centered on evaluating the match between the curriculum and the assessment, there has also been an increase in research aimed at developing methodology for assessing alignment. A review of the literature reveals that several alignment methods have been developed. According to Bhola et al. (2003), alignment methods can be categorized as low, moderate and high

complexity. The categorization of alignment methods is based on level of focus, that is, the number of dimensions considered in a particular study. For instance, a low complexity alignment study would only focus on the match between content of the items and the standards while a high complexity study would also consider other dimensions such as match in depth of content and the match between the levels of emphasis placed on a particular content area in the curriculum and on the assessment.

One implication of this categorization is that different alignment studies may come up with different results depending on the levels of focus employed. For example, Bhola et al. (2003) stated that an alignment study that does not consider the range of difficulty of items as a dimension may lead to misleading inferences about students' achievement and growth especially if students are to be classified into performance categories. As such, results from alignment studies of the same assessment but employing different levels of focus cannot be meaningfully compared. It is imperative then to develop methods for evaluating curriculum-assessment alignment that would produce results that are not dependent on the number of dimensions to allow for comparability of results over time or across states.

Almost all alignment methods reported in the literature involve SMEs. The SMEs are first trained to judge alignment against a specific set of criteria and decision rules (Ananda, 2003b). The SMEs are trained to ensure that they clearly understand the standards, the alignment criteria, and the scales being used to judge alignment. The content experts then review both the standards and the items to determine the match. Two issues need to be noted here. First, alignment methods currently in use almost entirely depend on human judgments about the match between the assessment and the curriculum. While expert judgments are essential in various steps in educational assessment, it is well

known that despite some training, humans may make errors of unknown magnitude in their judgment. With regards to alignment, Bhola et al. (2003) noted that SMEs may be overly generous in the number of matches that they envision. It was also observed in alignment studies in Nebraska that teachers worked harder to make sure that each item matched at least one content standard (Buckendahl, Impara, Plake, & Haack, 2001). Apart from the financial resources and time required to convene SMEs, having SMEs review each item and make judgments over multiple criteria can also be cognitively challenging. As Webb et al. (2005) noted, fatigued SMEs may not look closely to find the objective that matches a particular item but may choose a more familiar one. This would reduce the reliability of alignment results for some alignment criteria.

Second, the different alignment methods have different criteria and decision rules. Even in the cases where the criteria are the same, the operational definitions of those criteria vary from one method to another. Bhola et al. (2003) stated that "even in models of similar complexity that use the same labels for alignment criteria, alignment results depend critically upon the definitions of the criteria used" (p. 24). Thus appropriate interpretation of alignment results requires knowledge of the operational definitions of the criteria that define the model. For the sake of comparability and efficiency in terms of reduced costs, development or use of alignment methods that do not heavily depend on human judgment and apply a consistent set of criteria and decision rules is in order.

The other question that current alignment methods struggle with is: what constitutes sufficient alignment between standards and the assessment? As Ananda (2003b) noted, "…there is no hard and fast rule about what constitutes sufficient alignment" (p. 20). According to Ananda, one reason for the lack of such rules is "…when articulating expectations for what students should learn (what they should know

and be able to do), it is common for states to have different levels of statements, ranging from more global statements …to narrower more targeted statements clustered under the broader statement …"(p. 20). This means that choice of alignment method is partly dictated by the breadth of statements describing what students should learn. Consequently, results of an alignment study are dependent on the method. This outcome could pose problems in evaluating improvements in the assessment as measured by student achievement. Direct state-to-state comparisons could also be problematic.

Analysis of the various alignment methods reveals that some moderate and high complexity methods try to evaluate the agreement between the range of difficulty of the items on the assessment and the grade level of the students that the assessment is intended for. In this process, it is assumed that after some training the SMEs involved have a common understanding regarding the range of abilities of the students in the target grade. However, experience with other educational assessment processes that employ SMEs such as standard setting and content validity has shown that 100% agreement among SMEs is not always achieved. The magnitude of discrepancies among SMEs seems to increase with a decrease in quality of training. A good example is the 1990 Math standard setting for NAEP in which great variability was observed among SMEs in making item judgments despite training. The United States General Accounting Office (1993) claimed that the instruction given to the SMEs during training was not sufficient to bring the SMEs to a common understanding of what students at different achievement levels should know and be able to do. As a result each SME formulated their own definition of what a basic, proficient or advanced student can do resulting in large variability among SMEs in their judgments. The consequence of this variability was cut scores that were largely disputed and viewed as not representative of the knowledge and

skills of the students assessed. As Linn (1998) indicated, large "discrepancies of achievement levels and the location of the cut scores create a mismatch between what students with score in the range of the scale corresponding to a given achievement level are said to be able to do and what it is that they actually did on the assessment" (p. 20).

Other evidence that illustrates problems with SME judgments is found in an alignment study by Herman, Webb and Zuniga (2005), which found that while 20 SMEs with modest training had good agreement for item coding with respect to targeted topic and content, low reliability was observed for subgroups of the SMEs in their coding with respect to targeted objectives. The subgroups were groups of 6 raters (3 faculty and 3 teachers) drawn from the 20-member group and results from these subgroups were compared to results from the 20-member group. The 6 member subgroups had an 80% agreement with the 20-member group in terms of item content ratings. However, Herman et al. (2005) observed that "…the specific item and content on which they agreed upon varied across groups, suggesting that the 6-member groups tended to overestimate alignment …" (p. 28). These studies illustrate the point that some disagreement among judges should be expected due to differences in their bases for making judgments and their individual differences that may not be completely taken care of in training. High stakes decisions based on data from expert judgments should therefore be made realizing the weaknesses that are inherent in such data.

In addition to the fact that the SMEs may not have the same understanding of the students range of abilities, the other limitation is that the SMEs in alignment studies do not take into account the actual performance of the students. A mismatch between the SMEs' understanding of the range of student abilities at the target grade and what the

students can actually do could lead to alignment results that are erroneous and misleading.

Considering the issues raised above, it seems reasonable to consider other methods of evaluating alignment that would improve the utility of results. Desirable characteristics of such methods could be (a) accounting for student's actual performance on items, (b) reducing the reliance on subjective human judgment, c) applying consistent criteria for evaluating alignment, and d) producing results that are independent of the model applied in the alignment.

One method that could be used to evaluate the alignment of intended and actual item difficulty (range of difficulty) is item mapping. The next two sections briefly introduce item mapping method and how it could be applied to an evaluation of alignment.

1.3 Item Mapping

Webb (1999) defined alignment as "the degree to which expectations and assessment are in agreement and serve in conjunction with one another to guide the system towards student learning what they are supposed to know and do" (p. 4). This definition implies that the ultimate goal of alignment is to identify gaps in student learning through analysis of the correspondence between standards and the assessment. However, gaps in student learning can also be identified by determining what students know and are able to do.

One way of determining the knowledge and skills that students possess is to look at the actual student performance. Analysis of student performance could reveal their strengths and weaknesses and identify any shortfalls in the curriculum or instruction. Hence alignment could also take the form of matching the standards and what students

know and can do as evidenced by actual results of an assessment. In so doing, results of an alignment study would not only show the degree of agreement between the standards and the assessment, but also the match between the standards and actual student performance. Incorporating student performance into alignment would require a clear definition of "what students know and can do." Item mapping is one way that could be used to define what students know and can do.

Item mapping has been widely used in educational assessment in areas of standard setting (e.g., Wang, Wiser & Newman, 2001), scale anchoring (e. g., Gomez, Nash, Schedl, Wright, & Yolkut 2006), and score reporting (e.g., Kirsch, Jungeblut, Jenkins & Kolstad, 1993; Hambleton, 1997). Despite the various applications, the ultimate purpose of item mapping is to identify and describe what students at a specified level of achievement know and are able to do. For the purposes of this study, item mapping will simply be defined as the process of locating items along the test score scale. The idea behind item mapping is that given their characteristics, items could be systematically located on the test score scale based on some criteria. In most cases, the criterion used is the likelihood that examinees of a specified proficiency level have a high probability of success on the item.

One common approach for mapping items is the use of item response theory (IRT). IRT has been popular in most item mapping studies because in IRT models, student achievement levels and item difficulties are on the same scale. Thus, given an examinee's proficiency, items the examinee would most likely answer correctly can be identified. The phrase 'most likely answer correctly' is usually defined by the probability that the examinee gives a correct answer to an item. This probability is also referred to as

the response probability (RP) criterion in the literature. As it will be discussed later, choice of RP criterion has an impact on the results of item mapping.

In IRT models, each item is represented by an item characteristic curve (ICC), which gives the probability of passing an item for a given proficiency level. Figure 1.1 shows ICCs for three dichotomously scored items. The figure shows that item 3 has the lowest probability that an examinee would give a correct response throughout most of the score scale. This implies that item 3 is more difficult compared to items 1 and 2. Using a response probability of 70% (i.e., RP70), items 1, 2 and 3 would be mapped to scale scores of 300, 400, and 500 respectively. This means for example, that students with a scale score of 300 could be expected to correctly answer item 1 about 70% of the time. Similarly, students with scaled scores of 400 and 500 would be expected to correctly answer items 2 and 3, respectively, about 70% of the time.

Having located the items along the test score scale, the SMEs look at the items to identify and describe the knowledge and skills required for examinees along the score scale to give correct responses to the items. This task would be too demanding if descriptions for all the score points were to be written and if all items were to be used. To make the task more manageable, a handful of points along the score scale are chosen. These points, which are referred to as performance levels, are usually determined through a standard setting process. In some cases, a team of SME is convened and it is the team that decides how many performance levels will be reported for a particular assessment and also what labels would be used for each level.

Once the number and labels of performance levels have been agreed upon, performance category descriptions (PCDs) are developed. PCDs are detailed descriptions of the knowledge and skills that students reaching a particular performance level are

expected to demonstrate. The PCDs indicate the differences in accomplishment or mastery of students at different performance levels across the score scale. According to the National Research Council (NRC) (2005) determination of the number of performance levels and their descriptions should take into consideration the content being assessed in the test and inferences to be drawn from the scores.

The next step is identification of sets of items (also called exemplar items) that students at each performance level are very likely to answer correctly and that discriminate between performance levels. The exemplar items are used to develop performance descriptions for a particular score interval by describing the knowledge and skills required to successfully answer the items. The knowledge and skills in the performance descriptors are taken to represent the knowledge and skills that students at a specified performance level posses.

It is noted that performance levels and PCDs may be developed outside the standard setting process by a different group of SMEs. One example could be a study aimed at developing PCDs on the Scholastic Assessment Test (SAT) where initial cut scores were arbitrarily chosen by the researchers in consultation with policy makers (Hambleton & Sireci, 2008). Second, PCDs may be developed after exemplar items have already been identified.

1.4 Application of Item Mapping to Alignment

One way to make alignment studies more informative is to provide information that illustrates what students can do. This information would give an indication of how much students have learned and how much is yet to be learned. Item mapping could be used to provide such information. Item mapping can be applied to alignment in two ways.

First, item mapping could be used to describe what students at a particular grade can do. The first step in describing what students know and can do is choice of an RP value to be used as a criterion to distinguish students who possess the knowledge or have mastered a skill from those who have not. Given an RP criterion, IRT could be used to locate each test item on a proficiency scale. Given the proficiency range of students at a grade for which the test is intended, items that students are likely to answer correctly (given an RP value) can be identified. It is then concluded that students have mastered the skills required to successfully answer those items. Similarly, items that students in the proficiency range have a low probability of answering correctly would lead to the conclusion that students do not possess the knowledge and skills required to correctly answer those items. Further investigation would help to attribute the lack of knowledge to the curriculum or the instruction, and instruction can be targeted to provide the necessary knowledge and skills.

Item mapping could also be used in assessing vertical alignment. Although vertical alignment typically refers to equating tests across different grade levels, in this study, vertical alignment is defined as the process of mapping content standards of different grades or levels to a common scale. If content standards that span different grade levels can be located on a common scale, the progression of complexity of knowledge and skills across grades can be evaluated and students' progress along this progression of complexity can be tracked.

To place items from tests intended for different grades on the same scale, items that span grade levels must be calibrated onto a common scale, or tests designed for different grade levels must be vertically equated (Kolen, 2001). It is expected that items intended for lower grades would be at the lower end of the scale and items intended for

higher grades at the upper end. It is also expected that students at a particular grade have a higher probability (i.e., RP value) of success on the items intended for their grade compared to students in a lower grade on the same items (Thurstone, 1927). One assumption made here is that material taught at two different levels is different either in terms of content or at least cognitive complexity. Items mapping at an unintended grade level can then be looked at by content experts to discover why such "misalignment" occurred.

The literature available on alignment and item mapping is very limited. To the best of our knowledge, no studies exists that show how item mapping could be applied to alignment. A review of alignment studies that focused on the methods or application of the methods reveals that no alignment method so far incorporates student item responses. Thus, there is a need to extend the literature on both item mapping and alignment and explore the utility of item mapping in alignment. It is believed that the method introduced in the current study would not only reduce human involvement and hence error, but also enhance the utility of alignment studies by giving information about what students can do.

1.5 Purpose of Current Study

The purpose of this study is to investigate the utility of item mapping in evaluating the alignment between intended item difficulty (in terms of the grade span in which items are located) and actual item difficulty. Among other things, this study seeks to illustrate how item mapping could be used to assess alignment between curriculum and assessment. The study will also assess the impact of different RP values on curriclum-assessment alignment (i.e., how well items are located where they are expected). It is expected that greater alignment will be observed with a lower RP criterion than a larger

one. Lastly, the study will investigate the potential reasons for curriculum-assessment misalignment. The specific questions that this study intends to answer are as follows:

1. Can item mapping be used to assess the alignment between curriculum and assessment?
2. Do RP values have an impact on assessment- curriculum alignment results?
3. What are the reasons for assessment-curriculum misalignment?

From the above, it can be seen that the present study is unique in a number of ways. First, it introduces an efficient and convinient way of assessing alignment that employs limited involvement of SMEs. Second, it takes into account students' actual performance on a test to judge curriculum-assessment alignment. The current study therefore does not only extend the much needed literature on alignment and item mapping, it also introduces an innovative way of evaluating alignment.

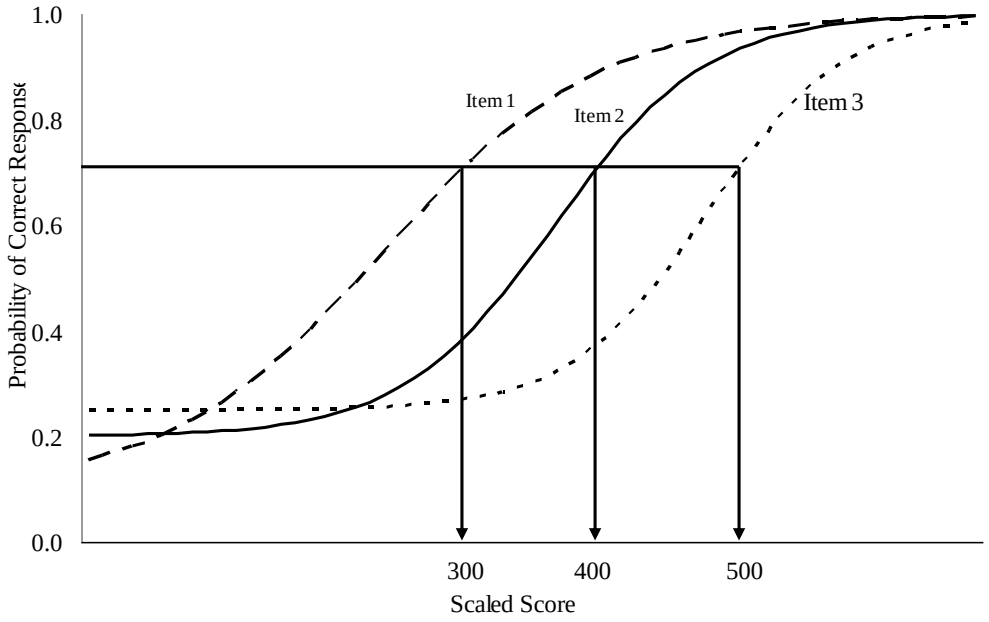


Figure 1.1. Item Characteristic Curves for 3 hypothetical items

CHAPTER 2

REVIEW OF LITERATURE

2.1 Overview

This chapter reviews literature on alignment and item mapping in general. The review begins with outlining why alignment is an important component in any educational system. This is followed by detailed a description of five alignment methodologies commonly found in the literature with a special focus on their strengths and limitations as well as differences among them. Examples of studies employing three of the methods will also be mentioned. The second part of the review describes item mapping methods available in the literature followed by literature on impact of choice of RP value on item mapping results. The review concludes with a brief discussion of item mapping as it has been applied to score reporting, scale anchoring, and standard setting with descriptions of studies to illustrate each application.

2.2 Alignment

2.2.1 Importance of Alignment

The main goal of alignment is to ensure that the standards, the instruction, and the curriculum are well coordinated to ensure student learning. When a test claims to measure achievement of some standards, it is important to evaluate how well the test represents those standards. This evaluation is important because if tests are not aligned to the standards, teachers are less likely to pay attention to the standards and this would affect the breadth of knowledge taught to students. Results of an alignment study therefore provide information on how well the assessment covers the curriculum as outlined in the standards and also gives insights into what is being taught in schools. Content gaps in the assessment or standards can then be determined (Ananda, 2003a) and

such information is important for policy makers to make informed decisions about the curriculum and the assessment.

Tindal (2005) adds that results of an alignment study may be used to identify areas where content standards may need to be clarified so that progression of knowledge across grades is more evident. Results of an alignment study may also be used in deciding whether restructuring of an assessment is necessary or not. If restructuring is necessary, alignment results would help to identify changes needed in the assessment.

Alignment also helps districts and states to compare their own standards and assessments to others (Ananda, 2003a). For example, a district may compare its results to state standards or a state may compare its results to standards of other states. This would help districts to evaluate their performance with respect to other districts or states. Ananda (2003b) also notes that alignment results could be used to provide evidence of content validity from an external source.

2.2.2 Alignment Models

The literature on alignment indicates that there are about five models that could be employed in an alignment study. According to Bhola et al. (2003), alignment models could be categorized as low, medium and high complexity. This categorization is based on the number of dimensions considered in a particular study.

Low complexity models look at alignment as the extent to which the content of items on a test match the content of the relevant standards. SMEs indicate the extent to which each item matches a content standard on a Likert scale. Moderate complexity models look at more dimensions other than simple content match. Examples of moderate complexity models include the Survey of Enacted Curriculum (SEC) and the Council for Basic Education (CBE) model. Lastly, high complexity models consider content match

and other dimensions such as cognitive complexity and performance match. Examples of high complexity models are the La Marca (2000), Webb (1997), and Achieve (2001) models. It should also be noted that although the goal of alignment is to ensure that the standards, the assessment, and instruction deliver a consistent message, not all models incorporate all the three components. Most models (as will become evident below) only consider alignment between the assessment and the standards. The SEC model is the only one so far that incorporates the instruction component.

The next sections give detailed descriptions of the CBE, SEC, La Marca, Webb, and Achieve models. Examples of studies employing the SEC, La Marca and Webb models are also described.

2.2.2.1 The Council for Basic Education Alignment Model

The Council for Basic Education (CBE) introduced a model with four dimensions: content, content balance, rigor, and item response type (Bhola et al., 2003). The content dimension looks at the match between content of the item and the standard. Content balance deals with distribution of the items assessing the standards while rigor relates to the match in cognitive complexity between the items and the standards. Item response type evaluates the appropriateness of the type of response being sought from the students in assessing the skill specified in the standards.

This model shares one weakness with other models (e.g., Achieve) in that it does not outline clear criteria for judging alignment.

2.2.2.2 The Survey of Enacted Curriculum (SEC) Alignment Model

The SEC is an example of a moderate complexity alignment method. Development of this model was motivated by the perceived need to develop "uniform descriptors of topic and categories of cognitive demand that together can describe the

content of instruction" (Porter, 2002, p. 4). One unique feature of the SEC methodology is that it does not only seek to establish alignment between curriculum (standards) and assessment, it also includes content of instruction into the picture. Thus, the SEC alignment model has content of the standards, assessment and instruction as its components.

The SEC model has two basic dimensions: content match and cognitive demand, which are assessed simultaneously by SMEs. These dimensions are used to create a two dimensional matrix with content on the horizontal and cognitive demand on the vertical axes. The content dimension lists the topics of the subject matter being assessed (e.g., linear equations and operations on polynomials in math) while the cognitive demand dimension lists categories of cognitive demand (Porter, 2002). The SEC model lists five categories of cognitive demand: memorize, perform procedures, communicate understanding, solve non-routine problems, and conjecture/generalize/prove. The matrix can then be applied to the assessment, standards or instruction. For the standards or assessments, SMEs identify the appropriate intersection between content and cognitive demand for each objective or item respectively. The resulting matrices (one for the assessment and one for standards) can then be compared to assess the degree of match and determine which areas are emphasized in one and not in the other.

Surveys are used to assess content of instruction. Using the same matrix described earlier, teachers code the instructional content based on amount of time spent on each topic (indicating coverage) and emphasis given to each category of cognitive demand. Each of the dimensions (i.e., coverage and emphasis) is coded on a 4-point scale. For coverage, 0 means not covered, 1 means slight coverage (less than one class or lesson), 2 means moderate coverage (one to five classes or lessons), and 3 means sustained

coverage (more than five classes or lessons). For emphasis, 0 means no emphasis, 1 means slight emphasis (less than 25% of time spent on topic), 2 means moderate emphasis (25-33% of time spent on topic) and 3 means sustained emphasis (more than 33% of time spent on topic). Again, the instruction matrix can be compared to the matrices for standards and assessment to judge alignment.

For the SEC model, alignment results can be summarized in two ways. First, an alignment index can be calculated to compare any two components (e.g., assessment and standards). The formula for the index is:

$$1 - \frac{\sum_{i=1}^{I} |X - Y|}{2} \qquad (2.1)$$

where X= assessment cell proportions, Y= standards cell proportions and I = number of cells in matrix (Porter, 2002). The proportions for X and Y would come from the SMEs ratings of both the assessment and the standards. The alignment index ranges from 0 (no alignment) to 1 (perfect alignment). Second, topographical maps can be created from the results to display the content that is emphasized by the assessment, the standards and the instruction. The maps can be compared to a) identify gaps among the three components for a particular district or state, and b) compare results across states or state to national results. Such a comparison is possible because as Porter (2002) noted, a common language is used to map the standards assessment and instruction.

The SEC model was used to assess alignment between content standards and assessments in math for the Goals 2000 project. Content standards and assessments for 7th grade math in four states were used in the study. In addition, content of the National Council of Teachers of Mathematics (NCTM) standard was also analyzed. The state standards and the NCTM standards were independently rated by 3 and 2 SMEs

respectively. Results of this study indicated that the assessments for each state were aligned to the states' standards as much as they were aligned to other states' standards. The average within state alignment index was 0.40 while the average between state alignment index was 0.39. Similar observations were made for the alignment between the states' assessments and NCTM standards where the average alignment index was 0.39 (Porter, 2002).

Blank, Porter, and Smithson (2001) carried out a study that illustrates use of the SEC model to measure assessment instruction alignment. Surveys were collected from 600 teachers in 20 schools across 6 states. The survey asked teachers to describe content of their instruction in 8^{th} grade math. The teacher's descriptions were compared to results of the content analyses of 8^{th} grade math assessments from the states and 8^{th} grade National Assessment of Educational Progress (NAEP) assessment. The results showed that state instruction was more aligned to NAEP assessment (average alignment index = 0.39) compared to within state assessment (average alignment index = 0.22). Between state alignment of instruction and assessment was slightly higher (average alignment index = 0.23) than within state alignment.

One advantage of the SEC alignment model is that results can be compared across states or district due to use of a common language to assess alignment among the standards, assessment, and instruction. Second, alignment results from the SEC model provide quantitative information about the alignment, which can be helpful in informing reform of the standards, assessment or both. Use of graphics provides an opportunity for visual presentation of alignment results which may be more appealing and easier for the general public to understand. Second, the visual presentation allows for comparison to find similarities and differences in content between the standards and the assessment.

One limitation of the SEC model pertains to how data about teacher's instructional practices are collected. As pointed out earlier, information about teacher's instructional practices is collected via surveys, and as Rothman, Slattery, Vranek and Resnick (2002) noted, this information may be prone to self report bias. In addition, teachers may not remember the details of their practices at the end of the year when this data is normally collected. Unlike the Webb model (discussed later in this chapter), the SEC does not provide criteria for judging sufficient alignment for some of the dimensions (Martone & Sireci, 2009).

2.2.2.3 The La Marca Model

One of the high complexity models was proposed by La Marca and his colleagues (2000). This model has content match, depth match, emphasis, performance match, and accessibility as its dimensions (Bhola et al., 2003). La Marca, et al. (2000) advocated for the evaluation of alignment between assessments and standards beyond simple content match arguing that "content match may be considered a necessary condition for an aligned system of assessments, but alone it is not sufficient to produce a high degree of alignment" (p. 18).

For the La Marca model, the content match dimension evaluates the agreement between content of the standards and assessment content. Depth match assesses the level of agreement between the cognitive complexity outlined in the standards and that reflected in the assessment. The emphasis dimension evaluates the agreement between the weight given to a particular content area in the assessment and the weight specified in the standards. La Marca et al. (2000) gave an example that a test that consists of a large number of computational problems but fewer problem solving ones is poorly aligned to standards emphasizing problem solving and reasoning.

Performance match deals with the agreement between what the students are asked to demonstrate in the assessment and the expected performance described in the standards. An aligned system will ensure there is a match between what is expected of the students and how it is reflected in the performance asked of the students in the assessment. For example, if students are allowed to use devices such as computers and calculators during instruction, such devices should be available during assessment for the two components to be aligned.

Lastly, accessibility seeks to establish if the range of knowledge required in the assessment matches the range of knowledge possessed by the students such that the assessment provides the opportunity for all students to demonstrate their level of proficiency. In other words, accessibility deals with issues of equity and fairness for students. According to La Marca et al. (2000), accessibility can be achieved if an assessment includes items that vary in difficulty to cover the different levels of achievement in a particular grade level. Thus an assessment should give an opportunity to all students to demonstrate their full range of knowledge and skills. Accessibility considerations are especially important if the assessment is also designed for use with student with disabilities and English language learners. If one assessment is administered to all students necessary steps should be taken to ensure that students with disabilities and English language learners participate in the test. Such steps may include accommodations like extra time, modified response type, large print, and modified question presentation format. Accommodations like these give the students an opportunity to display their level knowledge.

The major limitation of this model is that it does not give any guidance on how each of the dimensions could be evaluated. In other words, the model does not give clear

guidelines as to what level of agreement between the assessment and the standards is acceptable.

2.2.2.4 The Webb Alignment Model

Webb (1997) developed an alignment model with five categories: content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability. Each of the categories has some criteria for judging alignment. However, content focus is the only category that has been widely applied in most alignment studies applying the Webb model. As such, only brief descriptions of equity and fairness, pedagogical implications, and system applicability will be offered in this review. Greater details are included for the articulation across grades category because it directly relates to the type of data used in the study. The content focus category has six criteria for assessing alignment: depth of knowledge, categorical concurrence, range of knowledge, balance of representation, structure of knowledge, and dispositional consonance. However, only the first four have been widely applied in most alignment studies, so detailed descriptions of the four are given below.

Depth of knowledge correspondence evaluates the match in cognitive demands of the assessment versus the standards. In other words, depth of knowledge correspondence measures the degree to which the knowledge sought from students in the assessment has the same complexity as the knowledge the students are expected to have as specified in the standards (Tindal, 2005). The Webb alignment model identifies four levels of depth of knowledge from level 1 to level 4. The first level (recall) includes recall of facts, terms and definitions. Level 2 (skill/concept) items or objectives require students to engage in a mental process higher than mere recall of information (Webb et al., 1997). For example, tasks like comparing, classifying, and estimating involve students using information and

factual knowledge rather than just recalling the information. The third level which is called strategic thinking requires students to reason, develop a plan, and use evidence while level 4 (extended thinking) would engage students in complex reasoning and planning for a longer period of time.

According to Webb (2005), the depth of knowledge criterion is met if at least 50% of the items have a depth of knowledge level that matches the depth of knowledge of the objectives they assess. The 50% is based on the assumption that most assessments require students to correctly answer more than half the items on a test to pass (Martone & Sireci, 2009).

Categorical concurrence evaluates the extent to which the same or consistent categories of content appear in both the standards and the assessment (Tindal, 2005). An assessment would have high ratings for categorical concurrence if it includes items that target content from each of the broad categories in the standards. Webb (2002) suggested that six or more items should target each standard for an assessment to satisfy the categorical concurrence criterion. This number of items is based on the rationale that more items are required to make more reliable decisions regarding students' mastery of content.

The range of knowledge criterion assesses the degree to which the assessment covers the content dimensions represented in the standards. It measures the correspondence between the span or breadth of knowledge expected of the students and that required by the assessment. For example, if the standards require students to learn the order of operations in math, an assessment that only requires students to add would not satisfy the range of knowledge criterion. Range of knowledge criterion is met for a standard if the items targeting the standard are reasonably evenly distributed across the

objectives under the standard. Webb (1997) suggested that 50% of the objectives for a standard have at least one item targeting them for an assessment to satisfy the range of knowledge criterion. In other words the range of knowledge is satisfied if the assessment covers half of the domain.

The balance of representation criterion pertains to the distribution of items across objectives in the standards, that is, it assesses the extent to which the emphasis given to an objective on the assessment matches the emphasis in the standards. According to Webb (1997), objectives under a specific standard should be given relatively equal emphasis on the assessment. As such, items need to be evenly distributed across objectives for unbiased inferences to be drawn from the scores. Balance of representation is judged using a balance index which looks at the proportion of objectives assessed in the test relative to the number of items (Martone & Sireci, 2009). The formula for the balance index is:

$$1-\frac{\left(\sum_{k=1}^{O}\left|\frac{1}{O}-\frac{I_k}{H}\right|\right)}{2} \qquad (2.2)$$

where O =total number of objectives assessed for the subject, I_k = number of items corresponding to objective (k), and H = total number of items assessed for the subject domain.

Structure of knowledge comparability evaluates the match in the underlying conception of subject matter between the assessment and the standard (Webb, 1997). Dispositional consonance deals with the extent to which the assessment and the standards are in agreement in supporting broader visions of the learning the subject matter. Examples of the visions for the standards could be: develop positive attitude towards math and science (Webb, 1997).

The second category in Webb's model is articulation across grades and ages, which assesses the agreement between the standard and assessment on how they reflect student's growth and development over time. According to Webb (1997), assessments and standards should reflect the fact that students' understanding of concepts increases with their development. The extent to which the standards and assessment agree in the progression of knowledge across the developmental stages is a measure of articulation across grades and ages. According to Webb (1997), the two components can only be aligned if both are grounded in the same view of cognitive development that is backed by sound research.

Webb (1997) noted that research shows that "understanding is built gradually as new information is connected to existing networks of ideas" (p. 23). He argues that for strong alignment between the standards and assessment to exist, both should be based on this common view of how knowledge develops. In addition, the assessment and standards should reflect cumulative growth in content knowledge as students move from lower to upper grades. In other words, standards and assessments at a higher grade should require students to display more advanced skills and ideas compared assessments for to students at a lower grade. Both the cognitive soundness and cumulative growth in content knowledge components are evaluated at three levels: full, acceptable, and insufficient. Agreement between assessment and standards is full in terms of cognitive soundness if both are developmentally appropriate and show reasonable progression across grades. Similarly, an assessment and the standards are in full agreement in terms of cumulative growth of knowledge if both require students to display knowledge that matches with their cognitive development (Webb, 1997) and reflect the need for cumulative growth in content knowledge.

The equity and fairness category assesses how the assessment and the standards serve the full diversity of students in giving them the opportunity to reach the expectations and to demonstrate their knowledge. Webb (1997) cited social background and experiences, culture, race and gender differences as some of the factors that could result in assessment-curriculum misalignment.

The pedagogical implications category seeks to evaluate the consistency of the messages that teachers get from the assessment regarding practices in the classroom. Alignment is achieved if there is agreement among the standards, assessment and instruction practice. Lastly, system applicability category assesses the match between the standards and the assessment in terms of how realistic and manageable they are in the real world (Webb, 1997).

Webb (2006) applied his model to evaluate alignment of math standards and assessments for Wisconsin for grades 3 – 8 and grade 10. Eight reviewers (6 from Wisconsin and 2 from other states) participated in a three-day alignment analysis workshop. The reviewers consisted of math content experts, district math supervisors, math teachers, and math education doctoral graduate students. The alignment process began with training of reviewers. The reviewers were trained in the use of the four levels of depth of knowledge criterion by focusing on their definitions and examples. Then the whole group of reviewers was involved in determining the depth of knowledge of the objectives. This was followed by individual rating of the items. The depth of knowledge of the items was matched to the depth of knowledge of the objectives that the group had agreed upon. In this study, reviewers could match one item to up to three objectives. Reviewers could also make a note about any item that they felt exhibited inappropriate source of challenge.

A group review of the depth of knowledge of the standards showed that most of the objectives were at the skill and concept levels (i.e., levels 1 and 2). It was also observed that level 2 objectives increased across grades while level 3 objectives increased slightly. There were no level 4 objectives at any of the grades. Results also showed that alignment between standards and assessments was reasonable for four of the seven grades. Inadequate number of items assessing higher levels of depth of knowledge was the major reason for insufficient alignment for the other three grades. Based on this observation, Webb (2006) recommended replacement of lower level depth of knowledge items for the assessment to reach acceptable levels of alignment.

The Webb alignment model is a powerful tool that could be used to compare results across states. Comparison is possible because of the quantitative data that results from this model. However, alignment results from Webb's model can sometimes be misleading. For example, Martone and Sireci (2009) noted that an item that measures only part of a broadly stated objective is still considered to match the objective under Webb's alignment model. As such, results of the alignment can be inflated in as far as categorical concurrence, range of knowledge, and balance of representation are concerned.

2.2.2.5 The Achieve Alignment Model

The Achieve alignment model has six criteria: accuracy of the test blueprint, content centrality, performance centrality, challenge, balance, and range (Bhola et al. 2003). The process of alignment using the Achieve model follows three stages. The first is item by item analysis in which the items are compared to the standards to confirm the test blueprint, assess content centrality and evaluate performance centrality. The second

stage assesses challenge in terms of its source and level and the last stage assesses balance and range.

Confirmation of the test blueprint involves SMEs matching each item to the blueprint to ensure that every item in the assessment is related to at least one objective in the standards. The SMEs do this by way of discussion to reach a consensus about the degree of match between an item and the objective to which it is related. An item is considered to match an objective if it measures the same content specified in the objective (Rothman, Slattery & Vranek, 2002). In assessing the accuracy of the blueprint, the match between level of cognitive complexity required by the item and objective or the relative importance of the objective are not considered. Only those items that are matched to some objectives are considered for further analysis.

Content centrality evaluates the quality of match in content between the items and the standards. Each item is compared to the objective to which it is matched to evaluate the "specificity of the standard [Objective] and the extent to which the content to be assessed is evident from the reading of the item …." (Rothman et al., 2002; p.11). The degree of match is judged on a five-point scale where 0 means inconsistent match, 1A means a match where the degree of alignment is unclear, 1B means a somewhat consistent match as the item only measures part of a compound objective, 1C means a match where the objective is too specific to fully meet the item task, and 2 means a clearly consistent content match (Martone & Sireci, 2009).

Performance centrality seeks to establish the degree to which the cognitive demands of the assessment match the cognitive demands specified in the standards (Rothman et al., 2002). In judging performance centrality, SMEs scrutinize the action words in the item and the objective to see if the performance required in the item matches

the performance in the objective the item intends to measure. Each item can be matched to a maximum of two objectives where the objective that is most central to the item in terms of content is labeled as the 'primary match', while the other is labeled 'secondary match.' Judgment of performance centrality is made using the same rating scale as content centrality as described above.

In the Achieve model, the challenge criterion seeks to establish the level of mastery required for students to do well on a set of items (Rothman et al., 2002). Two factors are considered in evaluating challenge: source and level of challenge. Source of challenge evaluates if the difficulty in the item is related to some knowledge of content that students are required to posses or from other factors irrelevant to the construct being assessed. This is similar to Webb's challenge criterion in that both seek to assess if the item exhibits content that is not necessary for the examinee to correctly response to the item. In evaluating source of challenge, SMEs review the items to ensure that they are not flawed and the language level matches the grade level of the students. Each item is coded 1 if the source of challenge is appropriate and 0 if it is not (Martone & Sireci, 2009). On the other hand, level of challenge evaluates the range of difficulty of the items in relation to the student's grade level. To do this, SMEs first evaluate each item to establish the level of cognitive demand for each item. Based on the cognitive demand, each item is coded on levels 1 to 4 where Level 1 is recall or basic comprehension, Level 2 is application, Level 3 is strategic thinking, and Level 4 is extended analysis. Level of challenge is a qualitative decision that SMEs make after looking at a collection of items assessing a particular standard. SMEs make an overall evaluative judgment about how cognitively challenging the set of items is to students at a particular grade level relative to the standards.

The balance criterion evaluates match between the weight given to certain content in the assessment and the weight specified in the standards. According to Rothman and colleagues, the relative importance the assessment gives to the content skills should reflect that stated in the standards. SMEs evaluate balance by checking both the assessment and the standards to see if there are any objectives that are over-assessed or not assessed.

Bhola et al. (2003) stated that the range criterion seeks to evaluate the "degree to which an assessment contains items that measure knowledge and skills that are a representative sample of the content defined by the standards" (p. 24). To be representative the breadth and depth of the assessment should mirror the dimensions specified in the standards. In assessing balance SMEs evaluate the extent to which content areas deemed important in the standards receives the same emphasis in the assessment.

The range criterion in the Achieve model is similar to the range criterion in the Webb model. Range is a summative measure of the proportion of objectives assessing a standard that are measured by at least one item (Tindal, 2005). Ranges between 0.50 and 0.66 are considered acceptable, and ranges above 0.67 are considered good coverage (Martone & Sireci, 2009).

The Achieve alignment model was used to evaluate the standards and the assessments for Massachusetts grade 10 Math and English language arts (Achieve, 2001). In this study, the first step was to review Massachusetts' math standards. The standards were compared to standards for Arizona, Japan, and Achieves' standards because these were evaluated earlier to be among the best. The review was conducted by five national experts in standards. Two teams of reviewers (one for each subject) were convened to

asses the alignment of the assessments to the standards. The reviewer teams consisted of classroom teachers, curriculum specialists, and subject matter experts. The grade 10 math test for 2001 was aligned to standards for grades 9 – 10 in the 2000 math curriculum frameworks while the grade 10 English language arts test was aligned to grade 9 – 12 standards in the 1997 English language arts curriculum frameworks.

Results of the study indicated that the majority of items in the Grade 10 math test assessed content in the standards. Over 90% of the items were found to be aligned to content in the standards (Achieve, 2001). In terms of performance centrality, over 90% of the items were judged to seek the same performance specified in the standards. One math item was poorly rated because the standard it was intended to assess was stated in general terms posing problems for reviewers to determine direct alignment. The reviewers also found that the level of challenge for grade 10 math test was appropriate. There were very few occasions where items were flawed due to issues such as misleading graphics, multiple or no correct responses or ambiguous directions. However, reviewers pointed out that a large proportion (31%) of the test contained items that assessed grade 8 standards.

The grade 10 math test was found to contain items assessing all important aspects of the standards. Despite this finding, the balance of the test was judged to be uneven. The reviewers found that Algebra was overrepresented because items that the item writers thought measured Number Sense actually measured Algebra. Achieve (2001) recommended inclusion of more Number Sense items to balance the test. The authors also recommended to the state to consider mapping one item to up to two standards because they noted that many items assessed more than one concept.

Similar results were observed for the English language test where the items were found to measure only content in the standards. Content centrality results showed that 28 out of 34 items strongly or partially aligned to the standards. The rest of the items were partially aligned because the related content standards were not specific enough (Achieve, 2001). For the English test, 88% of the items showed high performance centrality in that the performance described in the standards matched the requirements of the items. The test also scored highly on the challenge criterion. About 25%, 65%, and 19% of the items were scored at recall, inference, and interpretation thinking levels respectively. No items were found to pose inappropriate source of challenge. The reviewers noted that the English language test required minor revisions in terms of balance. They recommended a balance of fiction and informational texts that appeared on the test. They also recommended mapping one item to multiple content standards and development of an item specific rubric for scoring writing assessments rather the generic rubric that was in place.

The availability of qualitative data from the Achieve model provides a thorough understanding of the degree of alignment. This information could be used to review the standards or the assessment. However, use of Achieve model requires a lot of time and skilled personnel, factors that could increase the cost of the study.

2.3 Similarities and Differences among Alignment Models

Considering the moderate and high complexity models described in this review, several similarities can be drawn across the models despite the differences in the number of dimensions in each model. First, all alignment models rely on SMEs to rate the degree of agreement between standards and assessment. The quality of alignment results is therefore somehow dependent on how well the SMEs understood the rating criteria

during training. In terms of assessing alignment, all models evaluate the match in content between the standards and the assessments. This helps to check that each item on the assessment measures content in some objective.

Second, the models also evaluate the extent to which the breadth of knowledge in the assessment reflects the breadth of knowledge in the standards. The five models all assess the degree of agreement between the cognitive demands specified in the standards and that required for examinees to give correct responses to items on the assessment. Although level of challenge is a very important aspect in alignment, all the five methods discussed in this review use SMEs to assess it. The current study is unique in that it represents a different way of looking at level of challenge by calculating item difficulty based on student performance, rather than relying on subjective judgment. Lastly, the models evaluate the relevance of the content on the assessment in measuring the content in the standards.

A number of important differences can be noted across alignment models. Some alignment models provide criteria for judging acceptable alignment (e.g., Webb and Achieve) while others do not (e.g., La Marca). The lack of criteria for judging acceptable alignment limits the utility of such models. Alignment models also differ in terms of the level of detail for matching standards to assessment. In some methods, matching is done at a more global level of the standards such as goals versus other models in which matching occurs at a much finer level such as the objective. The Webb model is the only model that can accommodate matching at any level of the standards. Such differences could have important implications on alignment results as well as on their comparability especially if the components being evaluated in the alignment study (e.g., assessment and standards) are written at different levels of detail.

Related to this issue is the observation that some methods provide both qualitative and quantitative alignment results (e.g., Webb, SEC, and Achieve) while others do not (e.g., CBE and La Marca). Both quantitative and qualitative results are important in comparing results across states and determining shortfalls in the assessment or curriculum. The other notable difference is that only the SEC alignment method incorporates instruction into alignment. This helps in providing information in the parts of the curriculum that teachers focus on.

2.4 Item Mapping

Item mapping has been used for three main purposes: score reporting, scale anchoring, and standard setting. In score reporting and scale anchoring, item mapping has mostly been used to identify items that could be used to describe the knowledge and skills that students at a specified proficiency level posses. In this sense, item mapping helps to make score scales and score reports more understandable to stakeholders. Item mapping has also been applied to the bookmark standard setting method to create ordered item booklets. This section discusses literature on item mapping focusing on available methods and the effect of choice of RP value on item mapping results. The section concludes with examples of studies that applied item mapping to score reporting, scale anchoring and standard setting.

2.4.1 Item Mapping Methods

Beaton and Allen (1992) described two methods of item mapping: the direct method and the smoothing method. These two methods are also commonly referred to as the empirical based and model based methods, respectively.

The direct method involves calculating the proportion of examinees answering an item correctly at different points on the score scale (Beaton & Allen, 1992). The first step

in the direct method is to create groups of examinees based on their scores. Examinees are categorized in such a way that all members of a group have score at or near an anchor point. Second, the proportion of students at or near the various anchor points that gave a correct response to an item is computed. The third step is to determine the items on which a high proportion of examinees in the first anchor point answered correctly. A 'high proportion' may be operationalized differently for different studies. For example, Zwick, Senturk, Wang, and Loomis (2001) defined it as 50%, 65%, and 74% of the examinees at an anchor point answering an item correctly. Fourth, items that high proportions of examinees at intermediate anchor points were able to answer correctly, but most of the examinees at the next lower level were not, must be identified. Finally, the groups of items identified for each anchor point are used to describe what examinees at a particular anchor point can do.

Beaton and Allen (1992) also described steps for item mapping using the smoothing (model-based) method as follows. First, choose a curve to represent the relationship between item responses and the score scale. The only requirement in this step is that the curve must be continuous and monotonically increasing. Second, fit the item characteristic curve to response data and locate the points at which a specified proportion of examinees can answer the item correctly. The third to fifth steps in this method are similar to those for the direct method described above.

The two methods of item mapping described above and their variations have been widely applied (e.g., Zwick, et al., 2001; Gomez, et al., 2007). For example, Zwick et al. (2001) employed a total of four variations of these methods in a study aimed at investigating item mapping methods. Two model-based (model interval and model midpoint) and two direct (empirical interval and empirical midpoint) methods were used.

The methods in each category differed in terms of how the probability of correct response was calculated. For the interval methods, the probability of correct response was calculated using responses from all examinees whose scores fell in a particular achievement level. On the other hand, midpoint methods used only those examinees within a specified interval around the midpoint of an achievement level (Zwick, et al., 2001). Two criteria were used to evaluate item mapping: RP values (R50, RP65, and RP74), and discrimination. Item discrimination was defined as the difference in item difficulty values between one achievement level and the next lower achievement level. Results based on the various methods across RP values and discrimination were compared to expert rating of the items.

Results showed that more exemplar items were identified when the discrimination criteria was disregarded. It was also found that the more exemplar items were identified using RP65 and RP74 compared to R50. Comparison across methods showed that model based methods matched more closely with expert's judgments than the empirical methods.

2.4.2 Response Probability and its Effect on Item Mapping

One decision that needs to be made in item mapping studies is how to define what level of student success in an item is adequate to indicate student's mastery of knowledge and skills assessed by the item. This level of success is what is termed response probability (RP). Response probability values are used to locate or map items along the score scale with the aim of describing the skills of examinees at specified score points. As the NRC (2005) stated, the decision about RP values is an important one because it affects interpretation of score levels. Various RP values such as 50, 65, and 80 have been proposed and used in item mapping studies (e.g., Kolstad, et al., 1998; Zwick, et al.,

2001). Both common sense and theoretical arguments have been put forward to justify use of particular RP values. For example Zwick and her colleagues justified use of RP50 arguing that "the 50% point marks the dividing line between cannot and can do" (p.16). A theoretical justification for R50 is based on the idea of item information as used in IRT. Based on IRT, the amount of information from an item is maximum when the probability of a correct response is 0.5 (assuming there is no guessing) (Kolstad, et al., 1998). Huynh (2006) noted that if p is the probability of a correct response, "the (total) item information for a Rasch and 2PL item is proportional to $p(1-p)$…" (p.20). This information is maximized at $p = 0.5$. The NRC (2005) stated that R50 could be defended statistically by considering the precision of estimated scaled scores. The authors noted that "the R50 values are always most precisely estimated…. The statistical uncertainty in the scale scores associated with RP values simply increases as the RP value increases above 0.50. It actually becomes very large for RP values of 90, 95, or 99 percent…" (p.85). Despite the support for R50, the study by Zwick et al. (2001) reported that SMEs indicated that 50% was insufficient to indicate student mastery.

Arguments for RP65 (or RP67) advance the idea that mastery of some skill would be evident if more students at a particular achievement level can do a task compared to those who cannot. Proponents for RP67 argue that if the number of examinees who give a correct response to an item is the same as those who do not (as is the case with RP50), it cannot be said that a substantial majority of students have mastered a skill. In other words, the idea of an examinee having a 50% chance of responding correctly does not connect well with the idea of mastery, hence the advocate for a larger RP value such as 67.

It is easier for stakeholders to associate mastery with RP67 because examinees are more likely than not to give a correct response to an item (NRC, 2005). Huynh (2006) provided a technical justification for use of RP67 by showing that for any dichotomous item the total information provided by the correct response is maximized if the RP value is greater than 0.50 for the one-, two-, and three-parameter logistic models. It is important to note that Huynh's argument clearly delineated total item information (which according to Huynh combines both the correct and incorrect response) from item information from the correct response. Huynh (2006) argued that under the Rasch and two- parameter logistic models, the item information from the correct response is given by $p(1-p)p$ which is maximized when $p = 0.67$. For the 3PLM this information is given by $p = (2+c)/3$, where c is the pseudo-guessing parameter. In terms of statistical precision, NRC (2005) pointed out that the error associated with estimated scale scores at RP67 was larger than at R50, but smaller than at RP80, hence RP67 is a good compromise between the two values.

Arguments for RP80 have also centered on the idea of a substantial majority of students being able to do a task. High RP values such as 80 are sometimes used when the type of decisions to be made based on the scores are high stakes in nature such that a lot of precision is required. A good example would be in certification and licensure exams where it is important to have a high degree of certainty that the certified or licensed individuals can perform the required task (NRC, 2005). However, some researchers have argued that the RP80 criterion appears to be too high (Kolstad, 1998). For example, RP80 was used to report results for the 1992 National Adult Literacy Survey (NALS). The results sparked a lot of debate to the extent that other stakeholders argued that use of

RP80 may have led to production of cut scores that were too high and so misrepresented the literacy levels of adults in the United States (NRC, 2005).

As Zwick et al. (2001) noted, the choice of RP criterion has an effect on item mapping results. Two studies can be used to illustrate this point. First, Kirsch et. al (1993) carried out a study to assess literacy levels among the adult population in America. The study employed item mapping to identify items that adults at specified proficiency levels could do. Employing a response probability criterion of 80%, the study found that 47% of American adults surveyed performed at the two lowest literacy levels (Kirsch, Jungeblut, Jenkins & Kolstad, 1993). When the data were reanalyzed using a response probability of 50%, only 20% performed at the two lowest literacy levels (Kolstad, 1996; NRC, 2005).

Zwick et al. (2001) carried out a study to investigate methods for item mapping. Using the National Assessment of Educational Progress (NAEP) data, the study employed RP values of 50%, 65% and 74% to identify items that could be used to exemplify skills and knowledge students at various achievement levels possessed. Results of the study showed that RP65 and RP74 yielded more exemplar items compared to R50.

NRC (2005) proposed three factors that could be considered in choosing response probability values for the purpose of standard setting. First is availability of empirical research about the effects of RP values on standard setting results. Such information is important in ensuring defensibility of cut scores. Second, NRC suggested use of statistical information about the precision of estimated scale scores for the various RP values. In general, amount of error associated with score estimates increases as RP values increase beyond 50%. However, NRC (2005) cautioned that much as R50 has the lowest error associates with estimated scores, some studies have shown that SMEs have

difficulties implementing this RP criterion. Third, choice of RP should take into consideration the objectives of the test, that is, the inferences drawn based on the scores and the consequences of those inferences. If test results are used for high stakes decisions such as licensure and certifying exams, a higher RP value might be considered.

2.4.3 Applications of Item Mapping

Item mapping has been applied to scale anchoring, score reporting, and standard setting. The following three sections describe studies that used item mapping for each of the three purposes.

2.4.3.1 Item Mapping and Scale Anchoring

Gomez et al. (2006) applied item mapping to a computer-based test for the purpose of scale anchoring. Their study involved the new Test of English as a Foreign Language internet based test (TOEFL iBT). The main goal of the study was to provide performance descriptors to test takers to help them correctly interpret their test performance. The descriptions would spell out typical proficiencies that were expected of examinees at each performance level. In this study, three performance levels were created by dividing the score scale into three equal percentiles; high, intermediate and low levels. An item was considered to map to the high or intermediate level if examinees at the specified level had an RP of 50%, the RP of examinees in the next lower level was less than 50%, and the differences in RP between the specified level and the next lower level was at least 20%. Items mapping to the low level were required to have an RP of 50% for examinees at that level. Once the items mapping to the different performance levels were identified, SMEs wrote descriptions of the knowledge, skills and abilities demonstrated by correct responses to the items (Gomez, et. al, 2006).

2.4.3.2 Item Mapping and Score Reporting

One study that illustrates the use of item mapping in score reporting is the NALS. This survey was aimed at assessing literacy levels of the adult population in America. In 1992, the NALS involved 26,000 adults aged 16 or older in 12 of the 50 states. Adults were assessed on their performance in three areas of literacy: prose, document, and quantitative. Results of the survey were also reported on three literacy scales, one for each literacy area. Item mapping was used to aid in interpretation of numerical scores representing adults' proficiency on the three scales (Kirsch, et al., 1993). For each item the point on the scale at which adults of some proficiency had an 80% probability of giving a correct response was identified. According to NRC (2005), RP80 may have been chosen because of the notion of mastery and to conform to item mapping for NAEP as this was the RP value used for NAEP at that time. Items mapping to each of the proficiency levels were then used to develop descriptions of the skills and knowledge that adults at that proficiency level demonstrated. Some items were selected and used as examples in the report (Kirsch, et al., 1993).

2.4.3.3 Item Mapping and Standard Setting

Wang (2003) described a study in which item mapping was applied to standard setting. Two features were unique to the study. First, an item map (described below) was presented to judges to help them make informed decisions about the items. Second, the study used the Rasch IRT model. Using this model, item difficulty and examinee ability are on the same scale and "when candidate ability equals item difficulty, then the probability of a correct answer to an item is 0.50" (Wang, 2003; p. 238).

The first step in the standard setting process was a discussion of the characteristics of the minimally competent candidate (MCC). Once a consensus was

reached on the definition of MCC, judges were presented with an item map in form of a histogram containing all items in the test arranged in columns according to their difficulty with each column representing a different difficulty. Items in each column were within two scaled score points from each other (e.g., 82 to 84). The columns were arranged from easy to hard where columns with easy items located to the left end of the graph and columns with hard items were on the opposite end. The standard setting facilitator then selected an easy item and asks the judges to make independent decisions about whether a MCC has a 50% chance of getting the item right. Then a more difficult item is selected and again judges are asked to make a decision on whether a MCC has a 50% chance of giving a correct response. This process continues until the judges reach a consensus that examinees have a 50% chance of giving correct responses to most items in some column. Since items in each column were within two scaled score points, the cut score was taken as the middle value of the level of difficulty of items in that column.

Results of the item mapping method were compared to standard setting results obtained using the Angoff method. Wang (2003) found that inter-judge consistency was higher for the item mapping method than for the Angoff method. Second, higher agreement amongst the judges was observed in the item mapping method than the Angoff method. Last, consistently lower cut scores were set using item mapping method than the Angoff method.

Another standard setting procedure that applies item mapping is the bookmark method (Mitzel, Lewis, Patz, & Green, 2001). In this method SMEs are provided with item booklets in which the items have been ordered from low to high based on their difficulty. The SMEs are then required to go through the booklet to find an item that a minimally competent examinee has less than the specified response probability of giving

a correct response to the item (Reckase, 2006a). Each SME places a bookmark in front of the item that they choose. The cut score is set to correspond to the difficulty of the item immediately before the bookmark or the average of difficulty of the items immediately before and after the bookmark. Studies evaluating the bookmark method show that the method generally results in lower cut scores than the Angoff method. For example, a study by Reckase (2006a) indicated that the bookmark method consistently underestimated cut scores. The study was based on the premise that each SME participating in a standard setting study would have an 'intended cut score (ICS)' that is based on their "internal understanding of the capabilities of examinees near the cut score" (p.5). The internal understanding of examinee capabilities is derived from their interpretation of policy and performance descriptors. Based on this premise, Reckase (2006a) postulated that the efficiency of a standard setting procedure could be evaluated on how well the ICS is recovered, the bias associated with the ICS estimates, and the standard deviation of the ICS estimates. Using simulation, Reckase (2006a) showed that the bookmark method tended to underestimate the ICS and had larger standard errors than the Angoff method.

Schulz (2006) defended the bookmark method stating that the standard setting evaluation framework proposed by Reckase and the simulation study lacked important details to explain the complexities that SMEs get into during standard setting. His main argument was that Reckase's study only focused on results of the first round of the bookmark method, a situation that misrepresents what happens in reality. Analysis of the Reckase- Schulz debate reveals that the empirical results used by Schulz in defense of the bookmark method were based on a variation of the bookmark called the Mapmark method (Reckase, 2006b; Schulz, 2006; Sireci et al., 2009).

2.5 Summary

This literature review has revealed that item mapping has been successfully implemented in various processes in educational assessment. Such areas include standard setting, scale anchoring and score reporting. In general, research shows that application of item mapping to these processes has been beneficial in production of results that are better understood by stakeholders and the public in general. For example, Ryan (2006) found that achievement performance level descriptions format was the most effective of the six score reporting strategies evaluated. Evidence also shows that improved standard setting methodologies that employ item mapping have gained popularity. This review has shown that curriculum-assessment alignment evaluation efforts have not fully tapped into the benefits of item mapping, particularly with respect to assessment of alignment in terms of cognitive complexity. Dimensions such as performance centrality (Achieve), cognitive demand (SEC), depth match (La Marca), and depth of knowledge correspondence (Webb) would benefit from item mapping. Additionally, item mapping would help in implementation and evaluation of Webb's articulation across grades dimension which aims at assessing agreement between assessment and standards in terms of their cognitive complexity progression across grade levels. Therefore, exploring how alignment would benefit from item mapping is an idea worth pursuing.

At present, there is no empirical evidence that links alignment results obtained from the various methods to actual student performance because none of the methods utilizes student item performance level data. Although evaluating the correspondence between alignment results from other methods and student performance is beyond the scope of this study, results of this study will provide a starting place as to how such an endeavor could be undertaken.

CHAPTER 3

METHODOLOGY

3.1 Overview

This study uses empirical data to illustrate use of item mapping in assessing alignment among curriculum, assessment, and instruction. A model based item mapping method was applied to the Massachusetts Adult Proficiency Test (MAPT) for Mathematics and Numeracy and for Reading. The MAPT for Math and MAPT for Reading tests are computerized multistage-adaptive tests. This chapter begins with a brief description of computer adaptive testing (CAT) focusing on multistage testing (MST) followed by a discussion of the MAPT score scale. The chapter also describes the data used for this study, how the data were analyzed and how results were summarized.

3.2 Computerized Adaptive Testing and Multistage Testing

Computerized-adaptive testing (CAT) refers to a system of test administration in which tests are administered using computers and adapted to an examinee's proficiency level. One methodology in CAT is known as multistage testing (MST). Multistage tests are those in which sets of items (called modules) that differ in difficulty are administered to examinees and examinees are routed to subsequent modules (stages) based on how they performed on the set of items (Hendrickson, 2007). Unlike in item-level CAT where adaptation occurs after every item, adaptation in MST occurs at the module level. That is, after an examinee responds to a set of test items (e.g., 5-10 items), an easier or more difficult set of items is selected for administration depending on how well they performed on the initial module. The first stage in MST is administration of a routing test. The aim of this test is to provide an initial estimation of an examinees' ability and based on this

estimate, a decision is made on which set of items (module) should be administered to the examinee in the second stage.

The MAPT for Math and MAPT for Reading are computerized multistage adaptive tests administered to adult learners with the aim of assessing their knowledge and skills in Mathematics and Reading respectively so that their progress in meeting educational goals can be evaluated (Sireci et al., 2008). The MAPT uses a six-stage test design (see Figure 3.1 & Figure 3.2). The test is organized in modules and panels. A panel is a collection of modules that defines all potential paths that examinees may be routed to when taking the test (Sireci, et al., 2008). Each of the MAPT tests consists of two panels. In MST, panels are analogous to alternate forms as defined in linear testing. The arrows in Figure 3.1 show some (but not all) potential paths that examinees may be routed to. Currently, there are no restrictions regarding the path that an examinee could take: that is, an examinee beginning the test at Beginning Basic may be routed to Low Adult Secondary based on their performance on the Beginning Basic items. The first time a student takes the MAPT s/he is randomly assigned to one of the two panels. The other panel is used for a second test administration. A total of 40 scored items are administered to each student across the six stages. Students take 15 items during the first stage and 5 items in each of the subsequent stages. Proficiency estimates at each stage are used to determine the set of items the examinee will take during the next stage. All items are dichotomously scored multiple-choice items with four answer choices. The next section briefly describes the MAPT score scale and how it was established.

3.3 MAPT Score Scale

Each panel of the MAPT tests is designed to assess students' proficiency in Math at four different educational levels. These educational levels are described by the United

States (US) Department of Education as part of the National Reporting System in adult education and are called Educational Functioning Levels (EFL). There are five EFLs assessed by the MAPT defined as Beginning Basic, Low Intermediate, High Intermediate, Low Adult Secondary, and High Adult Secondary. The US Department of Education's Office of Vocational and Adult Education (OVAE) established the National Reporting System (NRS) for Adult Basic Education (ABE), which requires states to measure ABE learners' educational gains as a core outcome measure of program effectiveness (Kaira & Sireci, 2007). All states receiving funds from OVAE must comply with the NRS requirements (see http://www.nrsweb.org/). These federal and state accountability demands were the primary factors motivating development of the MAPT.

For the MAPT, standard setting was used to determine cut scores that correspond to the NRS EFLs. Prior to standard setting, the NRS EFL descriptors were modified so that they are more appropriate for the MAPT. The standard setting process used the Item Descriptor Matching Method (IDM) to determine the cut scores for the EFLs (Sireci, et. al., 2008). The first step in the IDM procedure was a review of the EFL descriptions for the MAPT to have a clear picture of the knowledge and skills possessed by students in each EFL. The panelists then reviewed each item and matched the item to the EFL description that outlined the skills and knowledge required to correctly answer the item. Cut scores were determined using logistic regression of the difficulty parameter of each item within an EFL and panelists' classification of the items into each EFL (that is an item would get a 1 if it was classified in a particular EFL and 0 otherwise). The probability that a panelist would rate an item in a specific EFL was set at 0.50. IRT based cut scores were then transformed to the MAPT score scale that ranges from 200 to 700 with each 100-point interval corresponding to an EFL.

The same standard setting procedures were used to set cut scores for both Math and Reading tests. However, another standard setting study was carried out for the Reading test using the modified Angoff method. Results of the study showed there were some differences between the IDM and modified Angoff based cut scores. However, the differences were observed to be within measurement error expectations (Sireci, et. al., 2008). The cut scores for the MAPT for Math and Reading are shown in Table 3.1.

3.4 Data Source

Response data for both panels for the 2009 administrations of both Math and Reading was used in this study. About 7,361 examinees' responses to 362 Math items and 7,019 examinees' responses to 320 Reading items were analyzed. This study also utilized data on coding of the items into EFLs by item writers. The item writers specified Educational Functioning Level, content strand, and cognitive skill for each item. Later, an independent group of six SMEs who were involved in a content validity study also classified the items into EFLs independent of the item writers. The final classification for each item was taken to be the one that most (at least 4 or higher) SMEs that took part in the standard setting study agreed with. The distribution of the items across the EFLs based on the items writers' classifications is shown in Table 3.2. It is noted from the table that the Beginning Basic level for Math had the most items while the Low Adult Secondary level had the lowest number of items. For Reading Low Adult Secondary has the least number of items.

3.5 Parameter Estimation

The alignment method proposed in this study requires estimation of item and person statistics. IRT is one method that could be used to obtain these statistics from examinee response data. One advantage of using IRT is that examinee proficiency and

item parameters estimates are placed on the same scale such that given an examinee ability estimate, the probability of a correct response to an item can be determined. This study used parameter estimates from 2009 operational tests for both Math and Reading. The modified three-parameter logistic model (3PLM) was used to estimate examinee proficiency and item parameters. The mathematical form of the general 3PLM is given by

$$P(u_i = 1 | \theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - bi)}}, \quad (3.1)$$

where $P(u_i = 1 | \theta)$ is the probability of a correct response given examinee proficiency (θ), u_i is examinee response to item *i*, a_i is the item discrimination parameter for item *i*, b_i is the difficulty estimate for item *i*, c_i is the pseudo-guessing parameter for item *i*, and *D* and *e* are constants 1.7 and 2.718 respectively (Hambleton, Swaminathan & Rogers, 1991). A modification of the general model was used instead because it was noted that the data contained some very able examinees and some items had small samples. These two conditions could lead to problems in estimation of discrimination (a-) and pseudo-guessing (c-) parameters. To overcome this problem the a- and/ or c- parameters were fixed to 1.0 and/or 0.20 respectively, or a prior distribution was specified for some items (Sireci, et. al., 2008). The model for each item was determined through analysis and comparison of residual plots across various models. Parameter estimation was done using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996).

This study used a model based item mapping method and two RP values. These are described next.

3.6 Model Based Item Mapping Methods

Item mapping methodology was used to identify items that mapped to a particular EFL. The steps for the model-based item mapping method are described below.

(a) Obtain item parameter estimates for each item using the modified 3PLM. In this study, operational item parameter estimates for 2009 were used.

(b) Given the item parameter estimates, calculate the theta (θ) value required for an examinee to have some specified probability of correct response for each item. Two response probability values (.50 and .67) were used. Figure 3.3 illustrates the model-based item mapping method used in this study. As shown in the figure, the task is to find θ_1 and θ_2 for each item for which examinees have a probability of .50 and .67, respectively, of success on the item.

(c) Using theta values obtained in step (b) above and the cut scores, determine the EFL that each item maps to.

3.7 Response Probability Values

Two response probability values were used to determine the EFL to which each item maps - R50 and RP67. Two RP values are used to assess the impact of RP value on alignment. RP50 and RP67 have been chosen in particular for two reasons. First, these are the most common RP values in literature. Use of these RP values will therefore allow for comparison of the results of this study with findings of similar studies reported in literature. The second reason for the choice has to do with the purpose of the study. This study aims at illustrating how item mapping could be applied to evaluation of curriculum-assessment alignment. This requires some operational definition of what students can do. Based on literature, there seems to be a consensus that for tests that do not have very high stakes for individuals associated with their results, RP values higher than 67 may be too high. Considering the notion of defining examinee performance deemed satisfactory to warrant them being called knowledgeable, RP values lower than 50 seem to be less

defensible. For these reasons, it appears reasonable to use RP values of 50 and 67 as variables for the current study.

For the purpose of this study, an item was considered to map to a particular EFL if the probability of success in an item is .50 (for RP50) or .67 (for RP67) for examinees whose proficiency estimate (θ) falls within the specified EFL. Each item was considered to map to the lowest level where examinees have the specified criterion response probability of correct response or higher. For example, consider an item that examinees at the Low Intermediate level have a 53% chance of giving a correct response while the High Intermediate examinees have a 70% probability of correct response. This item was mapped to the Low intermediate level given the RP value of 50 but to High Intermediate given an RP value of 67.

After items have been mapped to the various EFLs, results were compared to the item writer/SMEs classification of the items. An item was considered to match or align to the intended EFL if the item mapping results agree with the item writer/SMEs classification. A situation where an item is mapped to an EFL other than the one the item writer/SME intended was considered a mismatch and misaligned.

3.8 Reasons for Curriculum-Assessment Misalignment

After items that do not map to the intended EFL were identified, SMEs (hereafter referred to as teachers) were convened for a one-day meeting to look at the items to find potential reasons to explain the misalignment. Due to resource constraints, this part of the study was only done for the MAPT for Math. Math was chosen because the researcher was more familiar with its content than with Reading. The teachers were drawn from current ABE teachers who have at least three years of experience teaching adult education students. This experience was required because of the need for the teachers to

have some knowledge of the ABE standards for Math and proficiency of ABE learners. Efforts were also made to ensure that teachers who teach students at all EFLs were included.

3.8.1 Procedure for the Meeting

The meeting began with self introductions of the participants followed by training that the facilitator conducted. The training sessions began with communicating the goal of the meeting, which was to review items that mapped to higher or lower EFLs than the SMEs had intended and suggest reasons for the misalignment. The teachers were then given a set of 6 items, which were used as practice items. The items were chosen in such a way that one-third were items that are "misaligned" with their intended level using the .50 RP value criterion and one-third that are misaligned using the .67 RP criterion. The other third comprised items that were well aligned. The well-aligned items were included in the practice set to serve as examples that teachers could draw upon. These items would help SMEs identify item characteristics that lead to examinee success and contrast those with the characteristics of the misaligned items.

The teachers looked at the items and the objective and level it was intended for and tried to find reasons why the item did not map to the intended level. Teachers were encouraged to look for such factors of the item as difficulty compared to proficiency of learners at a particular EFL, cognitive demand, language level, mathematical concept being assessed, and clarity. They were also encouraged to reflect upon their practice to determine if the topic assessed by the item was taught and how much it is emphasized.

The teachers first looked at the practice set of items individually followed by a group discussion. After the teachers had been trained, they were split into two groups. One group analyzed the items that failed to meet the RP50 criterion first, followed by

items failing to meet the RP67 criterion while the other group followed the opposite order. This was done to ensure that the order in which the items are reviewed does not have a significant impact on the results. The items were presented in two booklets with one booklet presenting items that misaligned at RP50 first and RP67 last while the second booklet had the opposite ordering of the items. Each teacher were presented with an item review sheet (see Appendix A) on which to record their reviews. Group discussions of some of the items followed individual review of the items. A questionnaire was administered to evaluate the item review process (see Appendix B). This questionnaire contained 5 Likert type and 2 open response items. The Likert type questions were rated on a 5-point scale from strongly agree to strongly disagree. In general, the survey sought teachers' views on aspects of the meeting such as adequacy of time for item review, adequacy of training and clarity of the item review task. The open-ended questions asked the teachers about some factors that they used in coming up with possible reasons for the observed misalignment and suggestions for the future.

3.9 Data Analyses

Both qualitative and quantitative analyses were carried out to summarize results for this study. Data collected from teachers on reasons for misalignment were analyzed qualitatively while data on item mapping were analyzed quantitatively. Descriptive and inferential statistics were used to summarize results.

Results were analyzed through comparisons to assess the degree of agreement between item mapping results and intended EFL for each item as indicated by SMEs. In this study, item classifications based on SMEs were regarded as "true" because two independent groups of SMEs vetted the item classifications. The comparisons were made at the item and content strand and cognitive skill level. The comparisons involved

examining the agreement between the model-based item mapping results and SMEs classifications for each RP value (R50 and RP67). Chi-square tests, correlations, and logistic regression were used to assess the degree of agreement. These are described next.

3.9.1 Chi-square Tests

The proportion of items mapped to the intended EFL by the respective RP value was calculated. Chi-square tests were used to determine if any observed differences in proportions across EFLs were statistically significant. This analysis took into consideration all the EFLs to which an item may potentially be mapped rather than a dichotomous analysis that only looks at whether or not an item maps to the intended level.

3.9.2 Correlation

For each RP value, the Spearman correlation between model-based item mapping results and SMEs' classifications of the items was calculated. The magnitude of the correlations was compared between RP values. Cohen's r^2 criteria (Cohen, 1988) were used to evaluate the magnitude of the correlations, where less than 0.10 is considered trivial, 0.10 -0.30 is considered small, 0.30 - 0.50 is considered moderate, and above 0.50 is considered large.

3.9.3 Logistic Regression

Logistic regression was used to explore the relationship between the criterion and independent variables and determine the amount of variance in the criterion variable explained by the independent variable. In this study, the criterion variable is whether or not an item maps to the intended EFL. The independent variable under investigation was the RP value. The logistic regression model for the study is given by

$$p(x=1 \mid EFL) = \frac{1}{1+e^{-(\beta_0+\beta_1 RP)}} \qquad (3.2)$$

where $p(x=1|EFL)$ is the conditional probability of an item being mapped to a specific EFL, β_0 is the intercept, and β_1 is the regression coefficient for the response probability criterion variable (RP). Change in Chi-square was used to determine amount of variance in the criterion variable accounted for by the independent variable and the full model.

3.10 Reasons for Misalignment

Reasons for misalignment were derived from written accounts provided by the teachers involved in this study. Content analysis (Borg, Gall & Borg, 1996) was used to analyze the data. The main objective of the analysis was to derive potential reasons that could be used to explain misalignment and shed more light on the characteristics of the items that contribute to misalignment. In conducting the analysis, reasons provided by the SMEs were coded into categories that pertain to characteristics of the item. The categories were as follows: cognitive complexity of the process required to respond to the item, difficulty of the item, language level of the item compared to level of students, clarity of the item, and emphasis placed on the topic during instruction. Further discussion of these categories is provided in the results section.

3.11 Analysis of Teachers' Survey Data

The Likert type responses from the survey were coded from 1 to 5 where 1 represented strongly disagree and 5 represented strongly agree. Descriptive statistics were used to summarize these data. Teacher's responses to open response items were analyzed qualitatively by identifying general themes.

Table 3.1. Cut Scores for the MAPT for Math and Reading

NRS EFL Boundary	IRT Cut score		
	Math	Reading	MAPT scale
Beginning Basic/Low Intermediate	**-0.23**	**-0.36**	**300**
Low Intermediate/High Intermediate	**0.43**	**0.84**	**400**
High Intermediate/Low Adult Secondary	**1.04**	**1.45**	**500**
Low Adult Secondary/High Adult Secondary	**1.74**	**2.05**	**600**

Table 3.2. Distribution of Math and Reading Items across ABE Educational Functioning Levels based on Item Writers' Classifications

Educational Functioning Level	Number of Math items analyzed	Number of Reading items analyzed
Beginning Basic	100	**62**
Low Intermediate	97	**121**
High Intermediate	94	**118**
Low Adult Secondary	71	**19**
Total	**362**	**320**

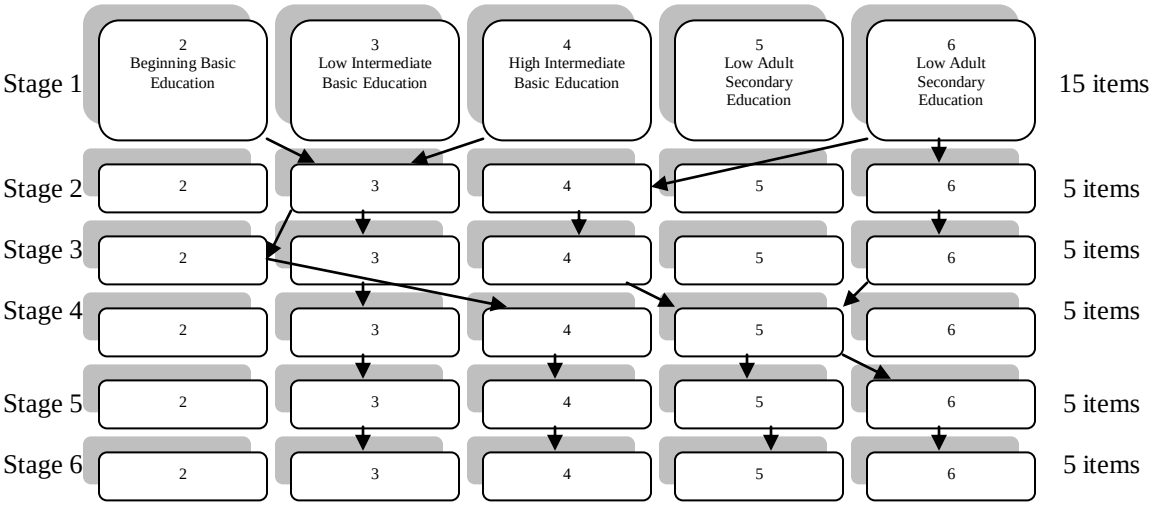
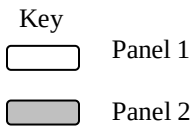


Figure 3.1 Multi-stage Test Structure for the MAPT for Math



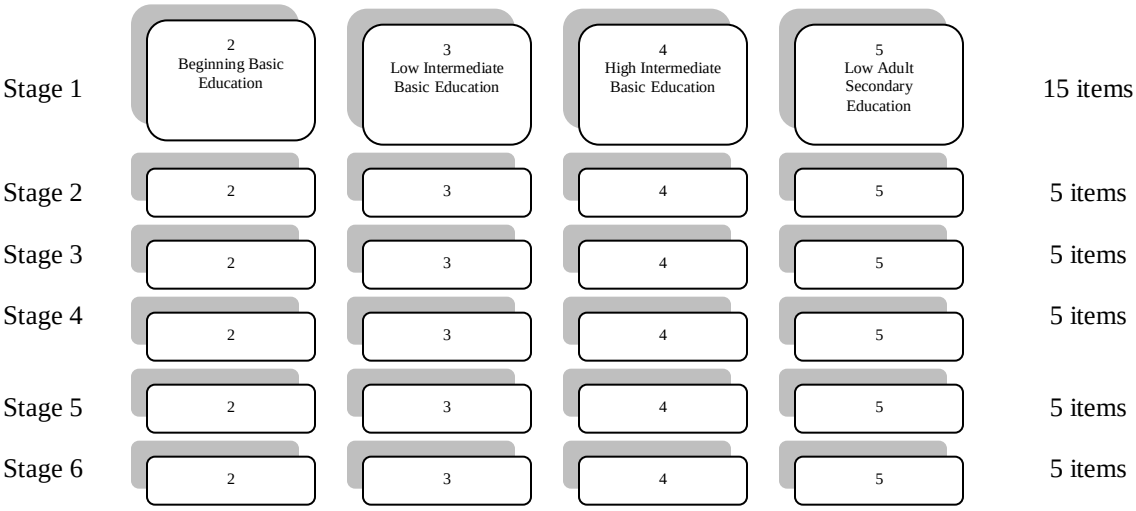
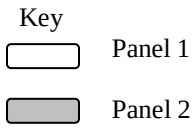


Figure 3.2 Multi-stage Test Structure for the MAPT for Reading



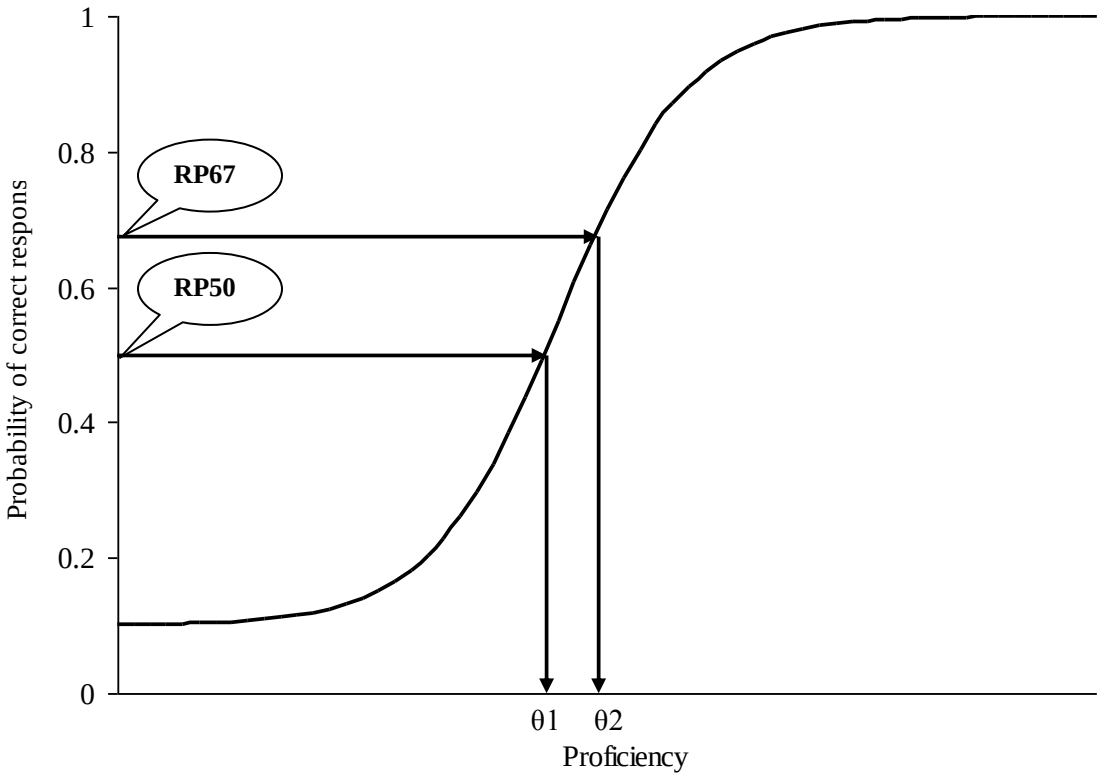


Figure 3.3. An Item Characteristic Curve Illustrating the Model Based Item Mapping Method

CHAPTER 4

RESULTS

4.1 Overview

This section presents results of the item mapping study. Math results are presented first followed by the results for Reading. For each subject, overall item mapping results are presented first followed by results stratified by content strand and cognitive skill level. Results of the study aimed at collecting reasons for misalignment are presented last.

It is worth noting at this point that IRT classification of the items used the 3PLM. This implies that each item was located on the proficiency scale based on the values of all the three item parameters (that is, discrimination, difficulty, and pseudo-guessing). This being the case, items will be located slightly lower than their *b*-value given RP50 as long as the value of the *c*-parameter is not close to zero. On the other hand, students will have a 67% chance of correctly responding to an item if their proficiency level is equal to the items *b*-value and the *c*-parameter is approximately equal to 0.35. If the c-parameter is less than 0.35, RP67 will always be higher than the b-value. Most of the items used in this study have c-parameter values less than 0.35 hence some over-classification is expected at RP67 compared to RP50.

Recall that in this study, the level to which each item is mapped using IRT is compared to the EFL for which the item is intended. Recall also that item writers classified each item to the EFL, cognitive skill area, and content area for which the item is intended. Later a second group of SMEs convened for the purpose of evaluating content validity classified the items again. Therefore for the purpose of this study,

intended EFL is defined as the EFL that test developers and/or SMEs (hereafter collectively referred to as SMEs) had finally indicated.

4.2 Mathematics

4.2.1 Overall Item Mapping Results

Table 4.1 shows the overall classification of the math items based on RP50 and RP67. It is interesting to note that for RP50, no items classified as Beginning Basic by the SMEs mapped to High Adult Secondary EFL. Similarly, no items that SMEs classified as High Intermediate or Low Adult Secondary mapped to Beginning Basic level. The chi-square for these results was 234.66 ($df = 15$, $p < .001$) implying statistically significant differences exist between the item mapping results and SMEs' EFL classification. The Spearman correlation between the two classifications is 0.692, which is considered moderate based on Cohen's r^2 criteria (r^2=.48).

Based on Table 4.1 the overall degree of exact agreement between SMEs' classification and IRT based item mapping at RP50 is 28.1%. This means that 28.1% of the items were mapped to the same exact levels as the SMEs classification. For RP50, the highest exact agreement between SMEs classification and item mapping results was at the Beginning Basic level where 34% of the items were mapped to the EFL intended by SMEs. For the Low Intermediate, High Intermediate and Low Adult Secondary levels, exact agreement was 28.9%, 28.7%, and 18.1% respectively. Considering items that mapped to adjacent levels, 47.1% of the items were mapped to one level lower or higher than the SMEs classification. Combining exact agreement and adjacent agreement as a measure of adjacent agreement between SMEs and item mapping classifications, it is observed that overall adjacent agreement at RP50 is 77.5%. The highest adjacent agreement was obtained at the Low Adult Secondary level. At this level, adjacent

agreement was 84.8%. This EFL also had the largest proportion of items (66.7%) that were mapped to the next higher level compared to all other EFLs. For the Beginning Basic, Low Intermediate and High Intermediate EFLs, adjacent agreement was 71%, 74.3% and 72.4% respectively.

Item mapping results based on RP67 are shown in Table 4.1. As expected, items are somehow over-classified for RP67 compared to RP50, that is, more items mapped to Low Adult Secondary EFL or higher and less items mapped to High Intermediate EFL or lower. Again, only 2 out of 100 items classified as Beginning Basic level by SMEs mapped to High Adult Secondary, while no items intended for High Adult Secondary level based on SMEs classification mapped to Beginning Basic and Low Intermediate levels. Based on the chi-square, statistically significant differences were observed between SMEs classification and item mapping results (χ^2_{15} =255.66, $p < 0.001$). The Spearman correlation between the two classifications was 0.71 ($r^2 = 0.50$), which is slightly higher than the correlation observed for RP50. From Table 4.1, the overall exact agreement between item mapping results and SMEs classification is 15.4% which is slightly above half the level of exact agreement for RP50. Results show that at RP67, the highest exact agreement between SMEs classification and item mapping is 20.8% at the Low Adult Secondary level. Exact agreement was 17%, 9.3%, and 16% for Beginning Basic, Low Intermediate, and High Intermediate levels respectively. At RP67 only 36.6% of the items mapped to one EFL lower or higher based on SMEs classification compared to 47.1% for RP50. Overall adjacent agreement for RP67 was 59.5%. The highest adjacent agreement between SMEs and IRT classification was 100% for the Low Adult Secondary level. Adjacent agreement was essentially the same for Beginning Basic and

Low Intermediate levels (50% and 50.5% respectively) and it was lowest for the High Intermediate level (47.9%).

In summary, it is noted that more congruence between item mapping results and SME classification of the math items was obtained at RP50. For both RP50 and RP67, the highest level of agreement as measured by adjacent agreement levels occurs at the Low Adult Secondary level. It is also interesting to note that for both RP values, larger proportions of items map to one EFL higher than the EFL for which the item is intended as classified by the SMEs. This may suggest that the items are generally harder than the SMEs had anticipated. Within RP50, adjacent agreement increased between Below Basic and Low Intermediate levels and also between High Intermediate and Low Adult Secondary level. The increase in adjacent agreement between the levels mentioned above was also observed for RP67. Adjacent agreement decreased between Low Intermediate and High Intermediate levels for both RP50 and RP67.

4.2.2 Item Mapping Results by Content Strand

The MAPT for Math is designed to assess learner's math skills in 4 content areas: Geometry and Measurement, Patterns, Functions and Algebra, Statistics and Probability and Number Sense. The content areas are hereafter referred to as Geometry, Patterns, Statistics, and Number Sense respectively. The distribution of the items across content areas is 84, 68, 93, and 116 for the four areas respectively. Overall item mapping results by content strand are presented in Table 4.2. It is evident from the table that regardless of content strand, RP50 tended to map more items to Beginning Basic, Low Intermediate and High Intermediate EFLs than RP67. On the other hand, RP67 tended to map more items to the Low and High Adult Secondary EFL compared to RP50.

Figure 4.1 shows item mapping results by content strand for RP50 and RP67 for the Beginning Basic level. The table shows that the highest level of exact agreement between SMEs classification and item mapping results was in Statistics where 44% of the items were classified to the same EFL by both. The lowest exact agreement (28.6%) was in Number Sense. It was also observed that adjacent agreement levels were relatively high across the content areas. For RP50, adjacent agreements for Geometry, Number Sense, Patterns, and Statistics were 62.5%, 71.5%, 75% and 76% respectively. Figure 4.1 also presents results for RP67 for the Beginning Basic level stratified by content strand. At this response probability, the highest exact agreement was in Geometry where 20.8% of the items were classified the same by SMEs and item mapping. The exact agreement for Number Sense, Patterns and Statistics content areas were 14.3%, 18.8%, and 16% respectively. The highest adjacent agreement at RP67 (64%) was obtained for Statistics content strand. Adjacent agreement for Geometry, Number Sense and Patterns content strands were 37.5%, 48.6% and 50.1% at RP67 respectively.

Item mapping results for RP50 and Low Intermediate level are shown in Figure 4.2. At this EFL, highest exact agreement was observed for Number Sense (35.3%) and the least was for Geometry (20.8%). At RP50, highest adjacent agreement was obtained for Statistics (81%) while the lowest was obtained for Patterns (56.7%). For Geometry and Number Sense strands, adjacent agreements were 76.6% and 79.4% respectively. Figure 4.2 also shows results obtained for RP67. It is interesting to note that none of the Patterns items intended for Low Intermediate level based on SMEs classification actually mapped to the Low Intermediate level. The highest exact agreement was in Geometry (16.7%) and exact agreement for Number Sense and Statistics were 11.8% and 4.8% respectively. As expected, adjacent agreement across content areas was less at RP67 than

RP50. The values were 41.7%, 64.7%, 33.3%, and 52.5% for Geometry, Number Sense, Patterns and Statistics respectively. It is interesting to note that at both response probabilities the lowest adjacent agreement was obtained in Patterns content strand. It is also noted that while the highest adjacent agreement was in Statistics at RP50, the highest adjacent agreement was observed in Number Sense at RP67.

Figure 4.3 presents item mapping results for RP50 and RP67 for the High Intermediate level. From Figure 4.3, it is seen that 28.7% of the items were classified as High Intermediate as intended by the SMEs. The highest exact agreement (40.7%) was in Number Sense while the lowest (22.2%) was in Statistics. Looking at RP50 results across content areas, adjacent agreement was the same for Number Sense and Statistics content areas (77.7%) while adjacent agreement for Geometry and Patterns were 57.2% and 79% respectively. Figure 4.3 also shows classification results at RP67 for the Low Intermediate EFL. More Statistics items (25.9%) were classified to the same EFL by both SMEs and item mapping compared to Geometry, Number Sense, and Patterns content areas which had exact agreement of 14.3%, 14.8% and 5.3% respectively. In terms of adjacent agreement, it is observed that the lowest was obtained for Geometry (33.3%) while the highest was for Statistics (59.2%). Adjacent agreements for Number Sense and Patterns content strands were 51.8% and 42.1% respectively.

Item mapping results for RP50 and RP67 for the Low Adult Secondary EFL are presented in Figure 4.4. As shown in the table, none of the items assessing Geometry, Number Sense, and Patterns content areas and intended for Low Adult Secondary mapped to Low Intermediate EFL at RP50. The exact agreements for Number Sense, Geometry and Statistics at RP50 were 20% each and the least agreement (13.3%) was

observed in the Patterns and Geometry content areas. At RP67, exact agreement ranged from 13.3% for Patterns to 25% for Statistics and Number Sense (see Figure 4.4).

In summary, results by math content strands show that the highest exact agreement was observed at the Beginning Basic EFL for the Statistics content area when using RP50 (44 %) and lowest was observed at the Low Intermediate EFL for the Patterns, Functions and Algebra content area using RP67 where none of the items was mapped as intended. In general adjacent agreement was higher for RP50 than for RP67 except at the Low Adult Secondary EFL where the two values were exactly the same.

4.2.3 Item Mapping Results by Cognitive Skill

The MAPT for Math analyzed in this study was composed of 114 items assessing learners' Knowledge and Comprehension skill, 175 items assessing Application skills, and 73 items assessing Analysis, Synthesis and Evaluation skills. For convenience, the three cognitive skill areas will be referred to as Comprehension, Application, and Evaluation respectively. Overall item mapping results stratified by cognitive skill are presented in Table 4.3. The table shows that for all cognitive skill levels, RP50 tended to map more items to the Beginning Basic, Low Intermediate, and High Intermediate levels compared to RP67 which tended to map more items at the Low and High Adult Secondary levels.

Item mapping results by cognitive skill for the Beginning Basic level are presented in Figure 4.5. The table shows that 46.5% of the items assessing Comprehension skills were mapped to the Beginning Basic level at RP50 compared to 30.2% at level RP67.Similarly, more Application (27.9 vs. 7.0%) and Evaluation (14.3 vs. 7.1%) items were mapped to the Beginning Basic level at RP50 than RP67. At RP50, adjacent agreement was 74.4%, 72.1% and 57.2% for Comprehension, Application, and

Evaluation cognitive skill areas respectively. Adjacent agreements at RP67 for the three cognitive skill areas respectively were 60.4%, 46.5% and 28.5%. In general fewer items intended to assess Comprehension skills at the Beginning Basic level were mapped to High Intermediate level or higher. Relatively more items intended for assessing Application and Evaluation skills mapped to High Intermediate level or higher at RP67.

Figure 4.6 presents results for the Low Intermediate level items stratified by cognitive skill. It is noted that for all cognitive skill areas, relatively few items intended for Low Intermediate level mapped to the Beginning Basic level. In fact, there were no items intended to assess Evaluation and Application skills that were mapped to the Beginning Basic at RP67. At RP50, more items (31.9%) intended to assess Application skills at the Low Intermediate level mapped to the intended level compared to Comprehension and Evaluation cognitive skill areas. The levels of adjacent agreement were 62.8%, 76.6%, and 70.6% for Comprehension, Application and Evaluation cognitive areas respectively. At RP67, the highest proportion of items mapped to intended level was from the Comprehension skill area where the exact agreement value was 12.6%. Adjacent agreement across cognitive skill areas for the Low Intermediate EFL was much lower at RP67 than RP50. The values were 57.5%, 51% and 35.3% for Comprehension, Application, and Evaluation cognitive areas respectively. For all the three cognitive skill areas and both response probability values, more items intended for the Low Intermediate level were mapped to the High Intermediate level compared to the proportion mapped to the EFL for which they were intended.

Results for the High Intermediate level stratified by cognitive skill area are presented in Figure 4.7. Overall exact agreement between item mapping and SMEs classification was 28.7% and 16% for RP50 and RP67 respectively. At RP50, exact

agreement was 36%, 24%, and 31.6% for Comprehension, Application, and Evaluation skill areas respectively. At RP67, the exact agreement for the three cognitive skill areas was much less (see Figure 4.7). Thus more items assessing Application skills at both RP50 and RP67 were misclassified than items assessing Comprehension and Evaluation skills. It is also interesting to note that neither RP50 nor RP67 mapped items intended for High Intermediate level to the Beginning Basic level. For Comprehension, Application, and Evaluation skill areas, adjacent agreement was 76%, 72% and 73.8% respectively at RP50, and 52%, 44% and 52.7% at RP67.

Figure 4.8 presents item mapping results for the Low Adult Secondary level by cognitive skill. Exact agreement for RP67 (21.1%) exceeded the exact agreement for RP50 (18.8%). Exact agreement for Comprehension and Evaluation cognitive skill areas were exactly the same for RP50 and RP67. For both response probability values, highest exact agreement was observed for Application cognitive area. The levels of adjacent agreement at RP50 were 100%, 94.3% and 100% for Comprehension, Application and Evaluation cognitive skill areas respectively. Similar to findings for the low and High Intermediate EFLs, more items intended for Low Adult Secondary EFL mapped to the adjacent higher EFL. None of the items intended to assess Knowledge and Evaluation skills mapped to the two lowest levels.

In general, results based on cognitive skill areas reveal that at both RP50 and RP67, more Evaluation items were misclassified compared to Comprehension and Application items. Another striking finding is that at the lowest EFL, greatest agreement between SMEs and item mapping classifications were obtained for Comprehension cognitive skills while at the highest EFL it was obtained for the Evaluation skill area.

Results also indicate that very few Application and Evaluation items intended for Low Intermediate EFL or higher mapped to the Beginning Basic level.

4.2.4 Logistic Regression

Logistic regression was conducted to explore the relationship between item mapping results and RP value. Results showed that RP value was a significant predictor of whether an item will map to intended EFL or not (χ_1^2 =17.318, p< .001). However, the results show that RP only accounted for 3.6% of the observed variance in item mapping results. It is also observed that using RP50, the likelihood that an item is classified as intended by SMEs is 2.142 times the likelihood at RP67.

4.3 Reading

4.3.1 Overall Item Mapping Results

Overall item mapping results for Reading for both RP50 and RP67 are shown in Table 4.4. The results show that in general differences exist between SMEs classification of the items and item mapping results. Based on the chi-square, these differences were statistically significant ($\chi_{12}^2 = 114.37$, p <.001). The Spearman correlation for the results at RP50 was 0.48, which is much lower than the correlation coefficient observed for the MAPT for Math at the same RP value. Based on Cohen's criteria (Cohen, 1988), the observed correlation is considered small (r^2=.23). From Table 4.4, it is seen that overall exact agreement between SMEs and item mapping classifications of the items is 45.3%. Overall adjacent agreement was 87.5%, which is 15% higher than the adjacent agreement obtained for Math.

Considering individual EFLs and RP50, it is noted that at RP50, considerably fewer items intended for the Beginning Basic EFL mapped to Low Adult Secondary EFL higher and fewer items intended for Low Intermediate or higher EFLs mapped to the

Beginning Basic level. It is also noted that the highest exact agreement was observed at the High Intermediate level. Exact agreement was 44.6%, 47.3%, 49.5%, and 30.2% for Beginning Basic, Low Intermediate, High Intermediate, and Low Adult Secondary EFLs respectively. Adjacent agreements for individual EFLs were 85.7% for Beginning Basic, 90% for Low Intermediate, 90.2% for High Intermediate, and 76.7% for Low Adult Secondary. Unlike the Math results where more items mapped to the next higher EFL than intended, less Reading items mapped to the adjacent higher level compared to the proportion of items mapped to intended level.

Table 4.4 also shows results for RP67. Similar to results for RP50, significant differences were observed between SMEs and item mapping classifications of the items ($\chi^2_{12} = 125.42$, p <.001). The observed Spearman correlation was 0.50 (r^2 = .25), which is essentially the same as that observed for RP50. Overall, there was a 22.8% agreement between the two sets of classifications indicating that most items did not map to the EFLs that SMEs had intended. Overall adjacent agreement at RP67 was 77.5%, which was 10% less than the value for RP50. Unlike results observed for RP50, more Reading items tended to map to the next higher level than the intended level at RP67. For example exact agreement for the Beginning Basic level was 16.1% while 41.1% of the items intended for this level mapped to Low Intermediate EFL (see Table 4.4).

The results summarized in Table 4 also show that exact agreement at individual EFLs was lower for RP67 than for RP50. For RP67, the highest exact agreement was observed at the Low Adult Secondary level where 34.1% were consistently classified by both SMEs and item mapping. Adjacent agreement was lowest for Beginning Basic (57.2%) and highest at the Low Adult Secondary EFL (90.9%).

Overall, the results for Reading are similar to those for Math. In both subjects, relatively higher agreement between item mapping results and SMEs' classifications was obtained at RP50 than at RP67. In addition, higher agreement was observed for the lower EFLs at RP50 and at higher EFLs at RP67. Conversely, the least adjacent agreement for RP50 and RP67 respectively were observed at the Low Adult Secondary and Beginning Basic EFLs.

4.3.2 Item Mapping Results by Content Strand

The MAPT for Reading was designed to assess learner's knowledge in three content areas: Comprehension, Vocabulary Meaning, and Word Recognition. The Reading assessment analyzed in this study was composed of 218 Comprehension items, 86 Vocabulary Meaning items, and 16 items assessing Word Recognition. It should be noted that the Word Recognition items were only developed for the Beginning Basic level. Overall, item mapping results by content strand for both response probability values are presented in Table 4.5. In general, more items are mapped to the High Intermediate and lower EFLs than the other two higher EFLs regardless of RP value.

Item mapping results by content strands for the Beginning Basic EFL are presented in Figure 4.9. The table shows that based on RP50, 58.3% of the Comprehension items at this EFL mapped to the intended EFL. Roughly similar proportions (35.3% and 33.3%) of the Vocabulary Meaning and Word Recognition items mapped as the SMEs had intended. The table also reveals that all the Word Recognition items that did not map to the Beginning Basic EFL mapped to the next higher EFL. Thus no Word Recognition items that SMEs classified as Beginning Basic mapped to EFLs higher than Low Intermediate. A large proportion (52.9%) of the items assessing learner's Vocabulary Meaning skills mapped to Low Intermediate rather than Beginning

Basic level for which they were intended. At RP50, adjacent agreement was 100% for Word Recognition, 75% and 88.2% for Comprehension and Vocabulary Meaning respectively.

Results based on RP67 show considerably lower proportions of items mapped to levels intended by SMEs. In general more items mapped to the next two higher EFLs compared to proportions that mapped to the Low Intermediate level. For example, a total of 94.1% of the vocabulary items mapped to Low and High Intermediate EFLs and 86.6% of the Word Recognition items mapped to the two levels. Note also that there were no Vocabulary Meaning and Word Recognition items intended for Beginning Basic level that mapped to Low or High Adult Secondary EFLs. Comparison between content strands shows that the highest adjacent agreement was obtained for Comprehension (62.5%). For Vocabulary Meaning and Word Recognition, adjacent agreement was 58.8% and 46.6% respectively.

Figure 4.10 presents results for both RP50 and RP67 for the Low Intermediate level. Results show that at RP50, a total of 45.5% of the Comprehension items and 52.5% of the Vocabulary Meaning items were consistently classified by both SMEs and item mapping. Lower proportions of items mapped to adjacent EFLs compared to proportions mapped to the intended EFL. Adjacent agreement for Comprehension and Vocabulary Meaning was 89.9% and 92.5% respectively. Only one item intended for the Low Intermediate level mapped to High Adult Secondary at RP50. Figure 4.10 shows that at RP67 exact agreement between SMEs and item mapping was less than half the agreement observed at RP50. The agreement was 14.8% and 20% respectively for Comprehension and Vocabulary Meaning respectively. As shown in Figure 4.10, the largest proportions of Comprehension and Vocabulary Meaning items intended for Low Intermediate EFL

mapped to the High Intermediate EFL. However, none of the Vocabulary Meaning items that SMEs classified as Low Intermediate mapped to Beginning Basic level. Only 2 (5%) Comprehension items for Low Intermediate level mapped to the next lower level. Adjacent agreement was 72.8% for Comprehension and 82.5% for Vocabulary Meaning content areas, which is lower than adjacent agreement for RP50.

Item mapping results by content strand for the High Intermediate level are shown in Figure 4.11. At RP50, about half (50.7%) of the Comprehension items SMEs classified as High Intermediate were mapped as intended. About 45% of the Vocabulary Meaning items were classified the same by SMEs and item mapping. Adjacent agreement was 88.7% for Comprehension and 95% for Vocabulary Meaning. At RP50, lower proportions of both Comprehension and Vocabulary Meaning items mapped 2 EFLs lower or higher than intended. Results for RP67 reveal that agreement between SMEs and item mapping for High Intermediate items was 32.4% and 20% for Comprehension and Vocabulary Meaning respectively. Similar to other findings in this study, larger proportions of Vocabulary Meaning and Comprehension items intended for the High Intermediate level mapped to the Low Adult Secondary EFL. Adjacent agreement fell at 83.1% and 90% for Comprehension and Vocabulary Meaning respectively.

Figure 4.12 shows that at RP50, greater classification agreement was observed for Vocabulary Meaning items than Comprehension items for the Low Adult Secondary level. One very striking observation is that larger proportions of Vocabulary Meaning and Comprehension items mapped to the next lower level than the proportions mapped to intended level. For example, 34.3% of the Comprehension items mapped to High Intermediate EFL versus 28.6% that mapped to Low Adult Secondary. Adjacent agreement was lower for Comprehension (73.3%) compared to Vocabulary Meaning

(88.8%). Results in Figure 4.12 show that exact classification agreement at RP67 was greater than the agreement at RP50. Similar to results at RP50, there was greater agreement in classification for Vocabulary Meaning than Comprehension items. At RP67, no items intended for Low Adult Secondary mapped to Beginning Basic EFL. Adjacent agreement was 88.3% for Comprehension and 100% for Vocabulary Meaning.

Results presented in this section generally show that at lower EFLs, greater exact agreement between SMEs classification and item mapping results was obtained for the Comprehension content strand. Lower levels of agreement were observed for the Vocabulary Meaning strand. However, adjacent agreement was higher for Vocabulary Meaning than for Comprehension at the lower EFLs. This means that most Vocabulary Meaning items intended for the lower EFLs mapped to the next higher EFL than the intended level. At the Low Adult Secondary EFL, the opposite observation was made. Greater agreement was obtained for the Vocabulary Meaning than the Comprehension strand and a large proportion of items mapped to the next lower level.

4.3.3 Item Mapping results by Cognitive skill

The MAPT for reading was composed of items assessing 3 cognitive skill areas: Locate/Recall, Integrate/Interpret, and Critique/Evaluate. There were 24 items assessing Critique/Evaluate skills, 169 items assessing Integrate/Interpret skills, and 127 items assessing learner's ability to Locate/Recall information. Figure 4.13 presents results by cognitive skill for Beginning Basic EFL. From Figure 4.13 it is seen that 48.9% of the items requiring students to Locate/Recall information mapped to the Beginning Basic level at RP50 compared to only 27.3% of the items assessing learner's Integrate/Interpret skills. This implies that 72.7% of the items assessing learners' Integrate/Interpret skills mapped to higher EFLs than the SMEs had intended. The proportion of Locate/Recall

items mapped to EFLs higher than Beginning Basic was much less. Results also show that at RP50, no items requiring students to Locate/Recall or Integrate/Interpret information that were intended for Beginning Basic EFL mapped to High Adult Secondary while 18.2% of the items requiring Integrate/Interpret skills mapped to Low Adult Secondary EFL. Adjacent agreements at RP50 were 91.1% and 63.7% for Locate/Recall and integrate/interpret items respectively. At RP67, none of the items intended for assessing learner Integrate/Interpret skills mapped to the Beginning Basic level, and only 20% of the Locate/Recall items mapped as intended. Adjacent agreement for Integrate/Interpret items was 45.5% compared to 60% for Locate/Recall items. These values are less than the values obtained for RP50. The observed low adjacent agreement levels imply that 40 to 55% of the items mapped to 2 EFLs higher than the Beginning Basic EFL for which the items were intended.

Results displayed in Figure 4.14 show that at RP50, more Locate/Recall items intended for Low Intermediate level mapped as intended. About 61% of the Locate/Recall items mapped as intended compared to 39.4% and 14.3% Integrate/Interpret and Critique/Evaluate items respectively that mapped as intended. A large proportion (57.1%) of the Critique/Evaluate items mapped to the High Intermediate level. Adjacent agreement was 91.1% for items requiring students to Locate/Recall, 90.9% for items requiring Integrate/Interpret skills, and 71.4% for items assessing Critique/Evaluate skills. Results obtained at RP67 were strikingly different from RP50 results. For example, none of the Critique/Evaluate items mapped to Low Intermediate EFL as intended by the SMEs, all critique/Evaluate items intended for Low Intermediate EFL mapped to Low and High Adult Secondary EFLs (see Figure 4.14). Exact agreement between SMEs and item mapping classifications were also low; 10.6% for

Integrate/Interpret skill area, and 25% for Locate/Recall skill area. Adjacent agreement ranged from 42.9% for Critique/Evaluate items to 82.1% for Locate/Recall cognitive skill areas.

Figure 4.15 presents results stratified by cognitive skill area for the High Intermediate EFL. The table shows that half of the items assessing Locate/Recall and Integrate/Interpret skills mapped to High Intermediate EFL as the SMEs intended. A total of 42.9% of the Critique/Evaluate items mapped as the SMEs intended. The other interesting finding is that at RP50, about 57% of the items assessing Critique/Evaluate skills mapped to the Low Intermediate EFL which is one EFL lower than the SMEs classification of the items. Adjacent agreement for each cognitive skill was high. The agreement was 100% for the Critique/Evaluate skill area, 92% for Integrate/Interpret area, and 81% for the Locate/Recall skill areas. Few items assessing Locate/Recall skills SMEs classified as High Intermediate mapped to Beginning Basic and High Adult Secondary EFLs.

Figure 4.15 also shows results obtained for RP67. Results show that the greatest exact agreement in classification was observed for the Integrate/Interpret cognitive skill area while the lowest was observed for the Critique/Evaluate area. Comparing results obtained at RP50 and RP67 reveals that greater agreement was obtained for Integrate/Interpret and Locate /Recall cognitive skill areas at RP50 than RP67. The table also shows that at RP67 about 52% and 59% of Integrate/Interpret and Locate/Recall items respectively mapped to the next higher level. Results indicate that adjacent agreement in classification was lowest for the Critique/Evaluate skill area (57.1%) and highest for Locate/Recall cognitive area (86.3%).

Item mapping results by cognitive skill for the Low Adult Secondary EFL are presented in Figure 4.16. It is observed that at RP50, classification agreement between SMEs and item mapping was 40%, 26.7% and 25% for Critique/Evaluate, Integrate/Interpret, and Locate/Recall cognitive areas respectively. About 40% of the items designed to assess Critique/Evaluate skills at Low Adult Secondary mapped to High Intermediate EFL. Adjacent agreement was 80%, 76.7%, and 75% for Critique/Evaluate, Integrate/Interpret and Locate/Recall skill areas respectively. Results obtained for RP67 show that exact agreement in item classification between SMEs and item mapping was 40%, 30% and 50% for Critique/Evaluate, Integrate/Interpret, and Locate/Recall cognitive areas respectively. Adjacent agreement was higher (90%, 89.6%, and 100% for Critique/Evaluate, Integrate/Interpret and Locate /Recall skill areas respectively) than adjacent agreement observed at RP50.

In summary, Reading results stratified by cognitive area show that greater exact agreement between SMEs and item mapping classifications was obtained at RP50 across all cognitive sill areas for the High Intermediate EFL and lower EFLs. For the Low Adult Secondary EFL, greater agreement was obtained at RP67 than RP50 except for the Critique/Evaluate cognitive skill where the two values were the same. Results also show that in general, higher adjacent agreement levels were obtained for Locate/Recall and Integrate/Interpret skills than for Critique/Evaluate skills.

4.3.4 Logistic Regression

Results of logistic regression of RP on item mapping results revealed that RP was a significant predictor (χ_1^2 =36.58, p < .001). The amount of variance in item mapping results that could be explained by RP value was 7.7% which is higher than the value

obtained for Math. Results also show that the likelihood that an item maps to intended EFL as determined by SMEs at RP50 was 2.8 times the likelihood at RP67.

4.4 Subject Matter Experts Study Results

A group of SMEs (hereafter referred to as teachers) was convened for a one-day meeting to look at the items that did not map as intended and suggest possible explanations for the observed misalignment. The teachers reviewed Math items only and this section presents results of that part of the study. The first section describes demographic characteristics of the teachers and the second discusses the characteristics of the items that were reviewed. Possible explanations for misalignment suggested by teachers are then presented. The section concludes with a summary of results of teachers' responses to the questionnaire.

4.4.1 Demographic Characteristics of the Teachers

A total of 7 teachers were involved in the study. The teachers came from all geographical locations across Massachusetts. Seventy-one percent of the teachers were female and the rest were males. As shown in Table 4.6, all teachers were Caucasian with teaching experience ranging from 3.5 to 32 years. All teachers had teaching certificates at elementary, high school, or adult education levels. The teachers employed in this study teach Math to ABE learners at various EFLs.

4.4.2 Items Reviewed by the Teachers

A total of 20 Math items were identified and selected for review. An additional six items were used as practice items. These items are presented in Table 4.7 with the practice items in bold. A review of all the misaligned math items at RP50 revealed that in general, misaligned items were slightly more discriminating and harder than the aligned items. The average discrimination and difficulty parameter estimates were 1.49 and 0.68

respectively for misaligned items versus 1.33 and -0.23 respectively for the aligned items. The average pseudo-guessing parameter estimate was 0.2 for both groups of items. This observation may imply that both the a- and b- parameters had an impact on alignment results.

Table 4.7 shows the item parameter estimates for each of the 26 items reviewed, the EFL each item each intended for, and the EFL the item mapped to at both RP50 and RP67. As is seen in Table 4.7, practice items were chosen in such a way that some items actually mapped to EFL intended by SMEs while other items mapped one to three EFLs higher than SMEs had intended. The table also shows that the items were well distributed in terms of the EFL for which they were intended. In addition, half of the items were chosen to represent misalignment at RP50 and the other half at RP67. The 24 misaligned math items selected for review had similar average discrimination parameter estimates to all misaligned items (1.51 vs. 1.49). However, the reviewed items were much harder ($\bar{b}$ =1.15 vs. 0.68) and had slightly lower average pseudo-guessing parameter estimates (0.17 vs. 0.20).

4.4.3 Possible Reasons for Misalignment

Teachers were employed to review misaligned items and suggest reasons for the misalignment. Six broad categories pertaining to characteristics of items were derived from the reasons provided by the teachers during the study. The categories were: item difficulty, cognitive demand of the item, language level of the item compared to language level of the students, the type of math concept being assessed, clarity of the item, and technical issues related to the item. Table 4.8 shows the number of items that the teachers thought exhibited issues related to each of the categories mentioned above. The categories are explained next.

It was observed that the math concept being assessed in the item was a factor contributing to misalignment. As shown in Table 4.8, this factor emerged as a reason for misalignment in 13 items. As the teachers noted, some mathematical concepts were generally harder for students. For example, teachers cited order of operations, conversion from one unit of measurement to another, finding the inverse, finding the circumference, and calculating the mean in a reverse order as some of the concepts that were challenging for students. Some teachers pointed out that students generally performed poorly on items involving the metric system of measurement because of the lack of experience with the system. As one teacher pointed out, students at the Beginning Basic EFL have problems converting millimeters to liters unless a conversion chart is provided. Multiplication and division by a fraction was also noted as one concept that was challenging to students. Another teacher noted that students at the Beginning Basic level generally confused symbols for less than and greater than and this could contribute to poor performance.

The teachers confirmed there were differences between the item writers' classifications of the items and item mapping results due to some characteristics that made the items easier than intended. This was observed in 3 of the 12 items in Table 4.8. One item had distractors that would be easily eliminated by even those students who did not have enough subject matter knowledge on the concept being assessed. For that item one teacher pointed out that the rest of the response options did not seem viable as possible answers except the correct response. Another teacher wrote "all the response options are clearly wrong and not even close to the correct answer." In other words the correct response was obvious enough to be easily spotted by less knowledgeable students. As such the difficulty of the item becomes much less than the item writer had intended.

Familiarity of the scenario presented in the item was another reason cited for lack of alignment in the second item. The item presented a scenario that may be familiar to examinees (that is, administration of prescription drugs to children) resulting in more examinees correctly responding to the item than expected. Item difficulty was reduced for third item because it required examinees to simply locate some points on a graph. In addition, there was a one-on-one correspondence between points on the graph and the response options making the item easier than expected. This easiness was hypothesized to occur because students either guessed the correct answer or chose the highest or lowest value, which happened to be the correct response.

The teachers also identified some aspects of the items that made 9 items more difficult. One aspect was the complexity of the numbers that students were required to manipulate. Teachers pointed out that some items tested at higher EFLs than intended because the numbers were too complex compared to the numerical ability of students at the EFL for which the item was intended. This could lead to students making calculation errors especially in situations where a calculator was not provided[1]. Item difficulty was also cited as a reason for item misalignment for 5 items that required multiple steps for students to arrive at a correct response. According to the teachers multiple steps increased difficulty of the items because of such factors as examinees skipping some steps or being unable to follow the steps in a correct order.

Drawing upon their experiences, teachers were also able to identify items that were too difficult for the EFL that the item writer had intended. One item that teachers identified as too difficult for students at a particular EFL presented stimulus material (i.e., drawing) that was too hard to interpret, while 3 items required students to derive new

[1] On the MAPT for Math, a pop-up calculator is available for some items that calculator availability is indicated by the item writer and required by the benchmark.

information given some facts (e.g., being able to figure out that the distance around an object is perimeter and be able to choose the correct formula, or be able to work with proportions to figure out the whole). Items requiring division by fractions, converting from fraction to decimal or from one unit of measurement to another were also seen to be difficult. Some SMEs also identified items requiring students to make generalizations as being generally harder. In addition, items that include viable distractors (e.g., median and mode in an item asking students to calculate the mean) also tested harder than expected.

The teachers identified cognitive demand of the item as a factor contributing to misalignment in 12 items (see Table 4.8). Items that require higher levels of thinking were generally harder. Most (9) items in this category asked students to derive and integrate new information into subsequent steps. These skills were cited as cognitively more demanding and hence more difficult for students. Most items involving multiple steps also fell in the category of cognitively more challenging. For instance, one item required students to first calculate the diameter given the radius and then use the diameter for other calculations. This question was hard for students at the EFL for which the question was intended. Another factor cited as increasing cognitive complexity of an item was presenting a geometry item without a diagram or providing a partially labeled diagram to examinees at lower EFLs. For example, in one item students at the Beginning Basic EFL were asked to calculate the perimeter of a rectangle without providing the visual of the rectangle. In another item, only two sides of a rectangular object were labeled in an item that asked students to find perimeter.

Based on their experience, the teachers noted that the two items were harder for Beginning Basic students because they were not yet able to derive a diagram from a description or be able to know that two sides of a rectangle are equal. An item that

required students to extrapolate was also cited as demanding higher levels of thinking. Tasks involving abstract thinking such as order of operations, or calculating the mean in a reverse order also tended to be more difficult. Most students at the Beginning Basic EFL had problems figuring out what the question was asking especially if it was not explicitly stated in the stem. For example, one item that asked students to find the distance around an object was found to be more challenging because, as teachers suggested, the word perimeter was not included in the stem. Similarly, teachers stated that students find it easier to solve a problem rather than ask them to identify the steps necessary to solve the problem.

Complexity of the language used in an item compared to reading level of the student was one factor that teachers suggested as contributing to misalignment in 11 items. Teachers noted that some items contained words that were hard for students at some EFLs and hence the poor performance on those items. For example, one teacher pointed out that reading and interpreting true/false statements was generally challenging for Beginning Basic students for whom English was a second language. This group of students also has problems with statements using passive voice. Teachers noted that vocabulary such as doubling every minute, consistent, mean, inequality, average, perimeter, more than half, three times more, twice as often, data, and equivalent were hard for students to comprehend especially at the lower EFLs. Students performed poorly on item that presented a scenario of bacteria in a culture to test students' ability to multiply. For this item, teachers felt that students performed poorly because they had problems with the phrase 'bacteria in a culture'. The teachers also noted that some items contained long and complex sentences that required more sophisticated reading skills that students for whom the item was intended did not possess. This led to students' lack of

understanding of the demands of the item and hence their poor performance. The teachers therefore recommended using shorter and simpler sentences so that language complexity does not affect student performance.

Eleven items were noted to exhibit some technical problems or ambiguities leading to students' poor performance. For example in one item, the stem did not state explicitly that students needed to provide their answer in different units of measurement than the units in the stem. Therefore students had to rework the question after realizing that the response options were in different units. In another question students were presented with a scenario where a fence needed to be put around a circular pond. However, the question did not specify that the fence also needed to be circular.

It was also observed that in three questions, the mathematical operators (such as plus, minus) were too small and some numbers were too close to each other and this could have resulted in students responding incorrectly. For two questions SMEs noted that the visuals (graphs) provided were confusing in that the lines were not very clear making it harder for examinees to identify the correct response. A similar observation was that three diagrams were poorly labeled. It was also observed that students could not tell the correct response to one of the items upon reading it without reading all the response options. This may have led the students to simply guess.

For 10 items, teachers cited lack of clarity of the item as a reason contributing to misalignment. For instance, one teacher noted that in one item, students needed to reformulate the question to be able to answer it because the question was not clear. For one question, teachers noted that the question was framed in such a way that it lead examinees to carry out a wrong mathematical operation. For this question, teachers recommended reordering the statements in the stem to improve its clarity. Teachers also

noted that presenting items in long sentences increased the likelihood of reducing the clarity of the item. As a result, the item became harder than intended. Similarly, teachers stated that some items contained information that was not necessary for students to respond to the item and that may have led to confusion among some students.

4.4.4 Teachers Responses to Questionnaire

A questionnaire was administered to the teachers to get their opinion about the meeting. The first five questions required teachers to indicate their level agreement to each statement on a 5-point scale ranging from strongly agree to strongly disagree. The first question asked teachers to indicate their level of agreement to the statement "the practice exercise helped me understand the item difficulty mapping task". All respondent either agreed or agreed strongly with this statement implying that the practice exercise helped to solidify teacher's understanding of the task. About 86% of the teachers disagreed strongly with the statement that inquired if they felt their opinions were ignored during discussion. The other 14% were neutral on this. All teachers involved in this study indicated that they clearly understood the task and that adequate time to review and comment on the difficulty of the items. Six of the 7 teachers disagreed with the statement that they needed more training to confidently complete the task. One teacher agreed with this statement.

Teacher were also asked about the factors they considered in reviewing the item to generate possible reasons for misalignment. The teachers cited language complexity, appropriateness of content for level of examinee, editorial errors in the question, and the number of steps required to solve the question. The teachers also mentioned the cognitive skill the item requires, the ability of examinees the item is intended for and also the vocabulary used in the item as some of the factors they took into consideration.

For future studies, the teachers suggested having two sections for feedback, one where they could note substantive issues pertaining to the items and the other where they could give editorial feedback.

Table 4.1. Math Overall Item Mapping Results for RP50 and RP67

Level based on SME	% items mapped to level based on RP									
	BB		LI		HI		LAS		HAS	
	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67
BB	34.0	17.0	37.0	33.0	24.0	31.0	5.0	17.0	0.0	2.0
LI	5.2	1.0	28.9	9.3	40.2	40.2	21.6	35.1	4.1	14.4
HI	0.0	0.0	11.7	3.2	28.7	16.0	33.0	28.7	26.6	52.1
LAS	0.0	0.0	2.8	0.0	16.7	2.8	18.1	20.8	62.5	76.4
Total	**10.7**	**5.0**	**21.5**	**24.0**	**28.1**	**24.0**	**19.3**	**25.6**	**20.4**	**33.1**

BB: Beginning Basic; LI: Low Intermediate; HI: High Intermediate; LAS: Low Adult Secondary; HAS: High Adult Secondary

Table 4.2. Math Overall Item Mapping Results by Content Strand

Level based on SME	% items mapped to level based on RP									
	BB		LI		HI		LAS		HAS	
	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67
BB	11.9	6.0	16.7	9.5	25.0	21.4	22.6	23.8	23.8	39.3
LI	9.5	4.3	25.9	14.7	33.6	29.3	15.5	26.7	15.5	25.0
HI	8.8	4.4	17.6	7.4	25.0	19.1	22.1	25.0	26.5	44.1
LAS	12.9	5.4	23.7	16.1	26.9	23.7	19.4	26.9	17.2	28.0
Total	**10.8**	**5.0**	**21.6**	**12.5**	**28.3**	**24.1**	**19.4**	**25.8**	**20.0**	**32.7**

BB: Beginning Basic; LI: Low Intermediate; HI: High Intermediate; LAS: Low Adult Secondary; HAS: High Adult Secondary

Table 4.3. Math Item Mapping Results by Cognitive Skill for all Levels

Cognitive Skill	% items mapped to level based on RP									
	BB		LI		HI		LAS		HAS	
	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67
Comp	21.9	12.3	19.3	14.9	26.3	24.6	17.5	25.4	14.9	22.8
Appl	6.9	1.7	24.0	13.7	28.6	24.6	20.0	26.3	20.6	33.7
Eval	2.7	1.4	19.2	5.5	30.1	21.9	20.5	24.7	27.3	46.6
Total	**10.8**	**5.0**	**21.5**	**12.4**	**28.2**	**24.0**	**19.3**	**25.7**	**41.2**	**32.2**

Comp: Comprehension; Appl: Application; Aval: Evaluation

Table 4.4. Reading Overall Item Mapping Results for RP50 and RP67

Level based on SME	% items mapped to level based on RP									
	BB		LI		HI		LAS		HAS	
	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67
BB	44.6	16.1	41.1	41.1	10.7	33.9	3.6	7.1	0.0	1.8
LI	7.0	1.6	47.3	16.3	35.7	57.4	9.3	20.2	.8	4.7
HI	3.3	2.2	23.1	4.4	49.5	29.7	17.6	52.7	6.6	11.0
LAS	9.3	0.0	14.0	9.1	37.2	27.3	30.2	34.1	9.3	29.5
Total	**12.8**	**4.1**	**34.7**	**16.3**	**35.3**	**41.3**	**13.4**	**29.1**	**3.8**	**9.4**

BB: Beginning Basic; LI: Low Intermediate; HI: High Intermediate; LAS: Low Adult Secondary; HAS: High Adult Secondary

Table 4.5. Reading Overall Item Mapping Results by Content Strand

Content Strand	% items mapped to level based on RP									
	BB		LI		HI		LAS		HAS	
	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67
Comp	12.4	4.6	29.8	12.4	38.5	39.4	14.7	32.6	4.6	11.0
V & M	10.5	1.2	41.9	23.3	33.7	44.2	11.6	25.6	2.3	5.8
WR	31.3	12.5	62.5	31.3	0.0	50.0	6.3	0.0	0.0	6.3
Total	**12.8**	**4.1**	**34.7**	**16.3**	**35.3**	**41.3**	**13.4**	**29.1**	**3.8**	**9.4**

Comp: Comprehension; V & M: Vocabulary and Meaning; WR: Word Recognition

Table 4.6. Demographic Characteristics of Teachers

SME	Sex	Race	ABE occupation	Years experience	Teaching certificate
1	Female	Caucasian	ABE teacher	7	Elementary special education certificate
2	Female	Caucasian	Pre GED teacher	10	K-8 certificate
3	Female	Caucasian	ABE math teacher	4	Secondary certification
4	Female	Caucasian	GED instructor	5	ABE certificate
5	Male	Caucasian	Pre-GED teacher	4	Principal/superintendent
6	Male	Caucasian	ABE teacher	3.5	English 7-12 licensure
7	Female	Caucasian	ABE teacher	32	Certificate in English

Table 4.7. Misaligned Math Items Reviewed by Teachers

Item	Item parameters			Level written to	Level mapped to (RP50)	Level mapped to (RP67)
	a	b	c			
1	1.55	1.12	0.20	2	5	6
2	1.46	0.66	0.14	2	5	5
3	2.08	0.84	0.26	2	5	5
4	1.04	0.77	0.13	2	5	6
5	**1.55**	**0.69**	**0.20**	**2**	**5**	**5**
6	**1.47**	**-1.26**	**0.21**	**2**	**2**	**2**
7	1.31	0.89	0.20	3	5	6
8	1.53	1.47	0.04	3	6	6
9	1.64	1.30	0.17	3	6	6
10	1.43	1.62	0.13	3	6	7
11	1.55	1.26	0.20	3	6	6
12	**2.34**	**0.97**	**0.17**	**3**	**5**	**6**
13	**1.54**	**-0.88**	**0.21**	**3**	**2**	**3**
14	1.10	1.90	0.23	4	6	7
15	0.89	1.63	0.20	4	6	7
16	1.06	1.48	0.14	4	6	7
17	1.50	2.16	0.20	4	7	7
18	3.08	1.23	0.20	4	6	6
19	**1.42**	**1.79**	**0.21**	**4**	**6**	**7**
20	**1.56**	**0.10**	**0.20**	**4**	**4**	**4**
21	0.67	-0.37	0.21	5	3	4
22	1.84	1.82	0.10	5	7	7
23	1.55	2.14	0.00	5	7	7
24	1.40	-0.15	0.15	5	3	4
25	0.89	1.85	0.21	5	6	7
26	1.83	1.56	0.05	5	6	7

Table 4.8. Summary of Reasons for Misalignment

Reason	No. of items
Math concept assessed	13
Item difficulty	12
Cognitive demand	12
Language level	11
Technical issues with item	11
Item clarity	10

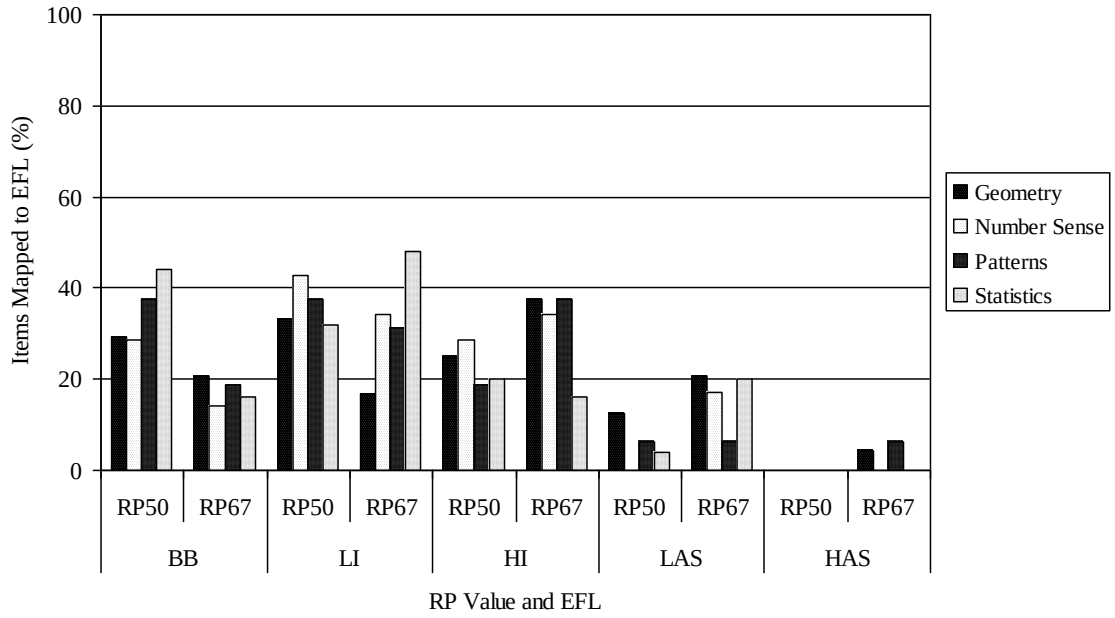


Figure 4.1. Math Results by Content Strand for Beginning Basic EFL

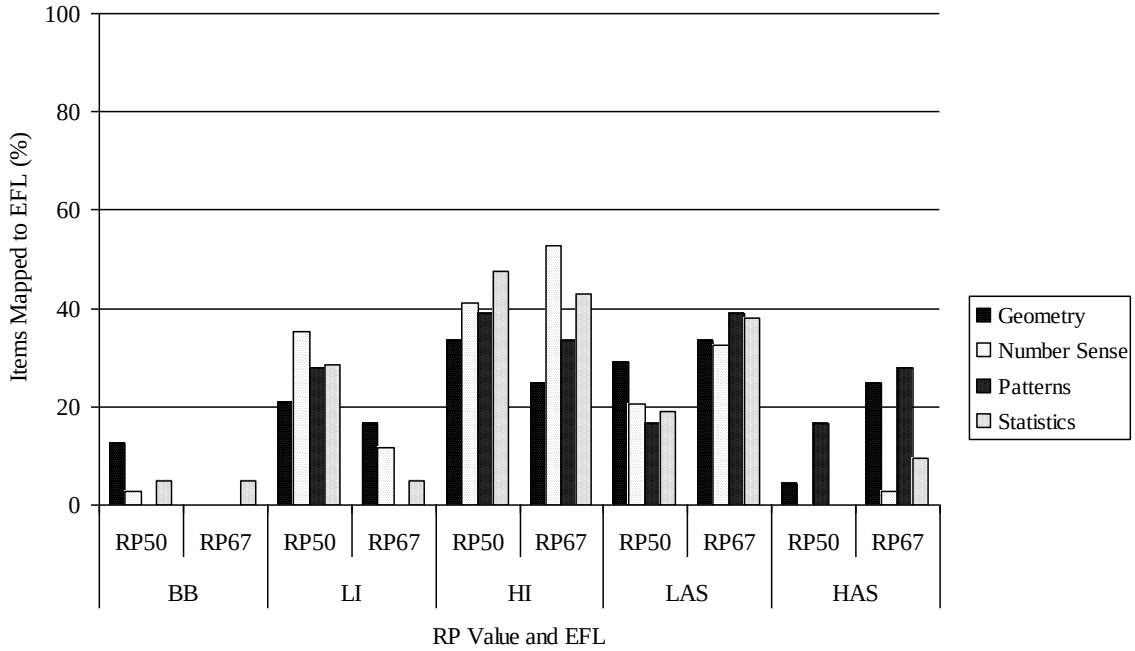


Figure 4.2. Math Results by Content strand for Low Intermediate EFL

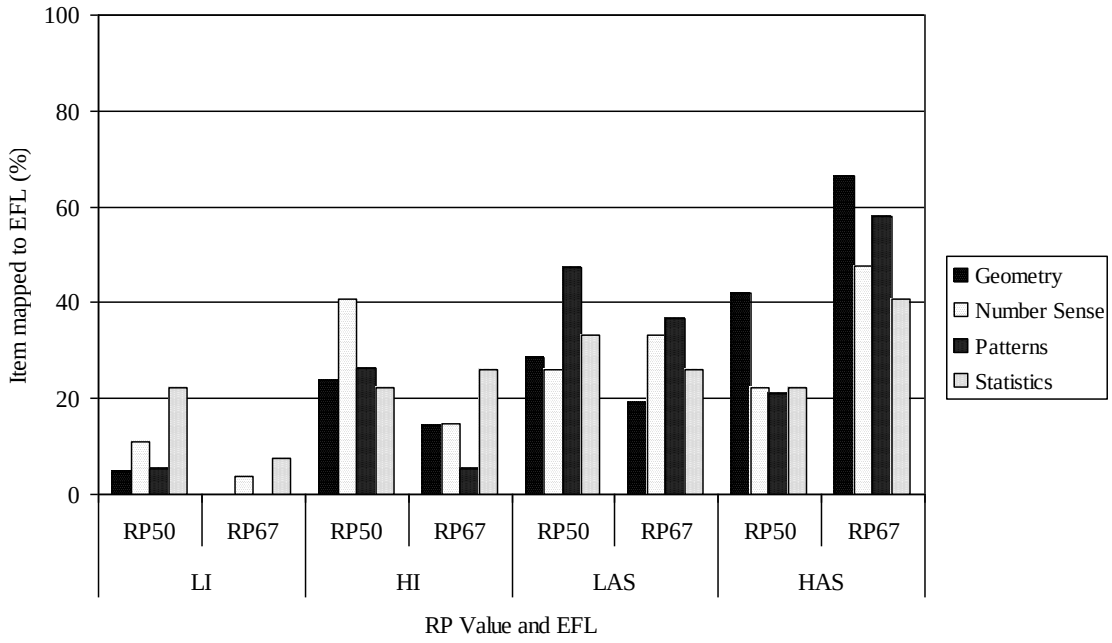


Figure 4.3. Math Results by Content Strand for High Intermediate EFL

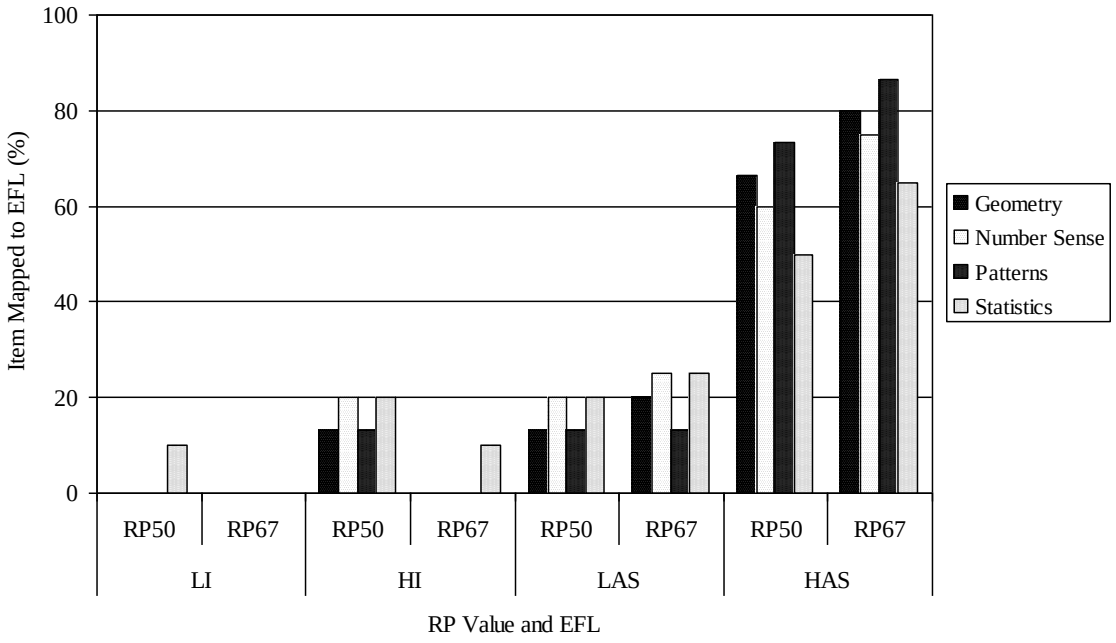


Figure 4.4. Math Results by Content strand for Low Adult Secondary EFL

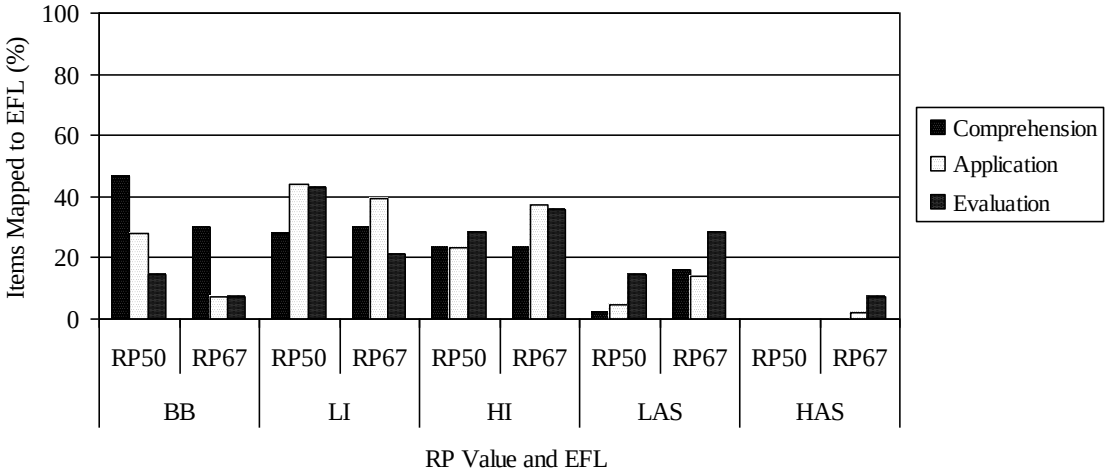


Figure 4.5. Math Results by Cognitive Skill Area for Beginning Basic EFL

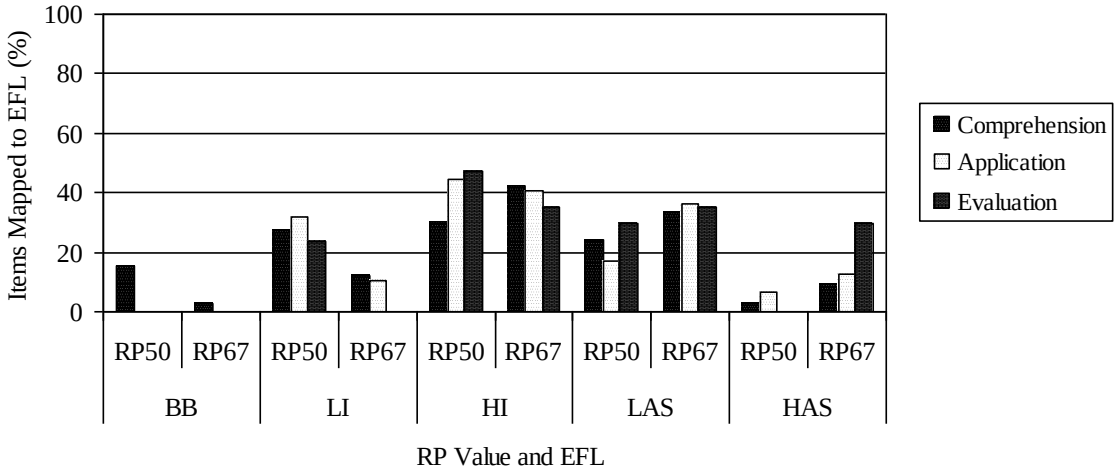


Figure 4.6. Math Results by Cognitive Skill Area for Low Intermediate EFL

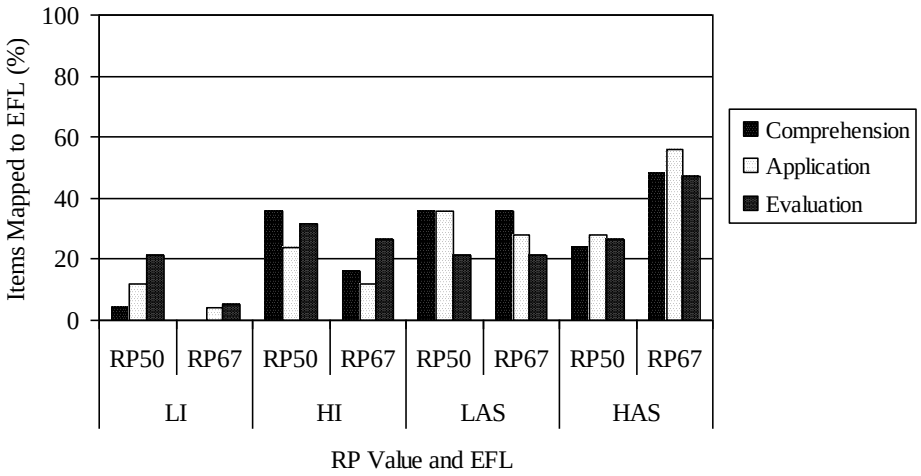


Figure 4.7. Math Results by Cognitive Skill Area for High Intermediate EFL

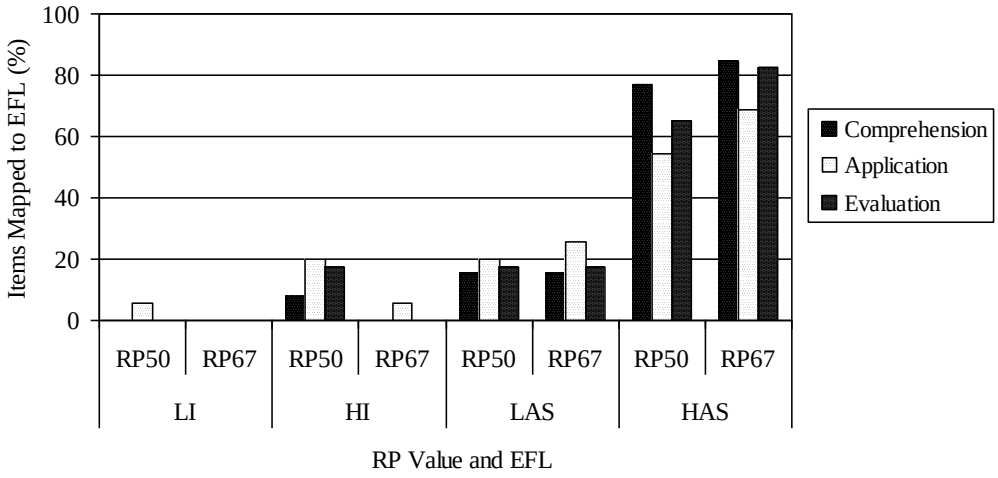


Figure 4.8. Math Results by Cognitive Skill Area for Low Adult Secondary EFL

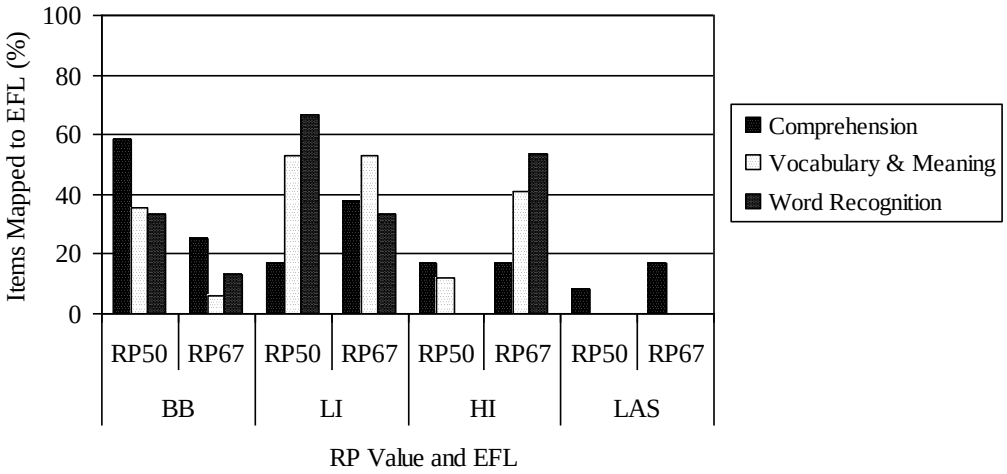


Figure 4.9. Reading Results by Content Strand for Beginning Basic EFL

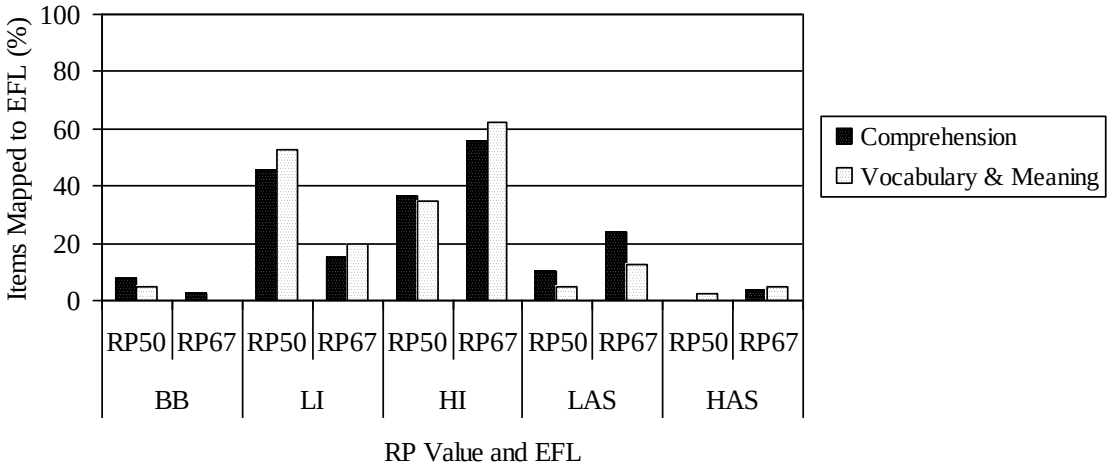


Figure 4.10. Reading Results by Content Strand for Low Intermediate EFL

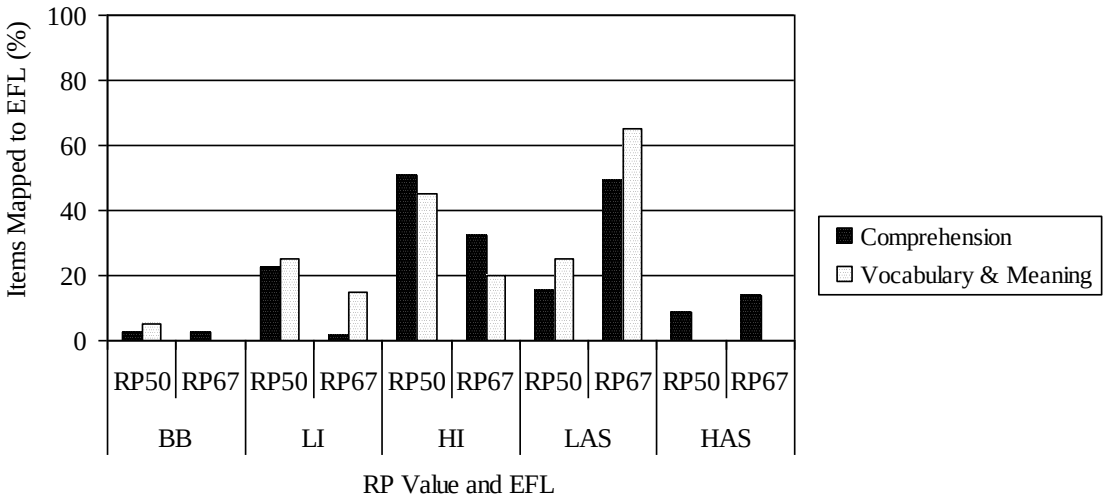


Figure 4.11. Reading Results by Content Strand for High Intermediate EFL

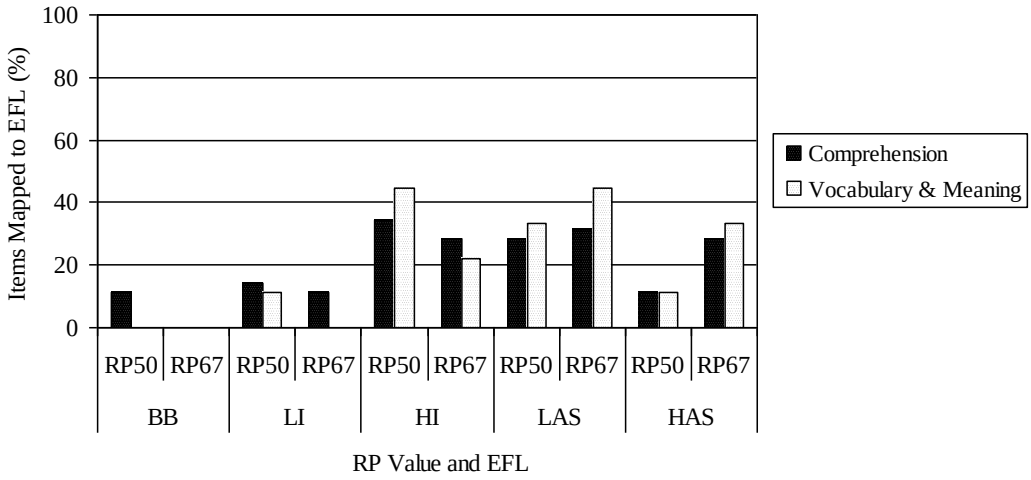


Figure 4.12. Reading Results by Content Strand for Low Adult Secondary EFL

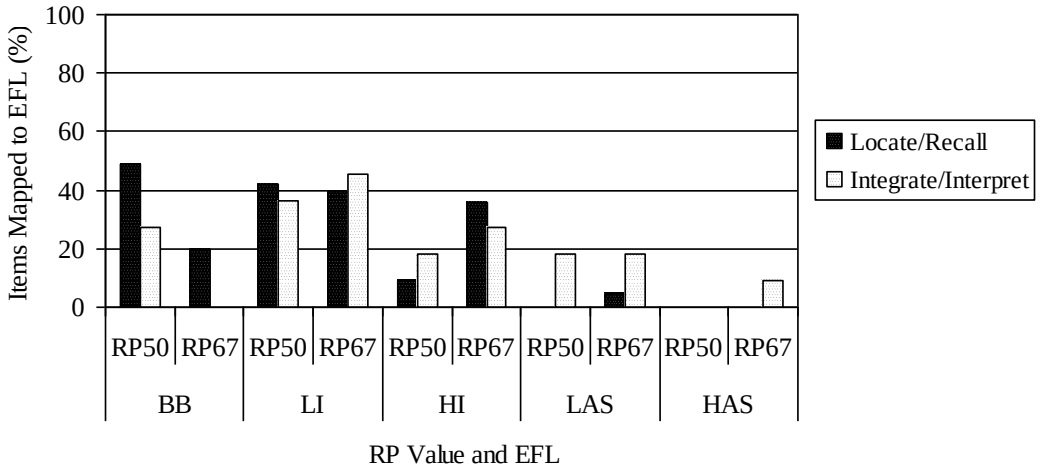


Figure 4.13. Reading Results by Cognitive Skill for Beginning Basic EFL

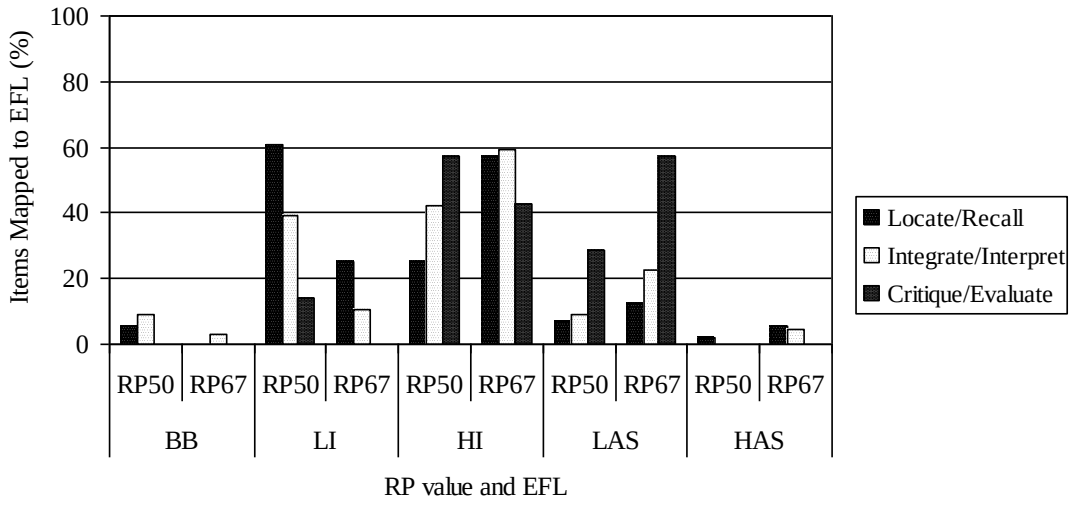


Figure 4.14. Reading Results by Cognitive Skill for Low Intermediate EFL

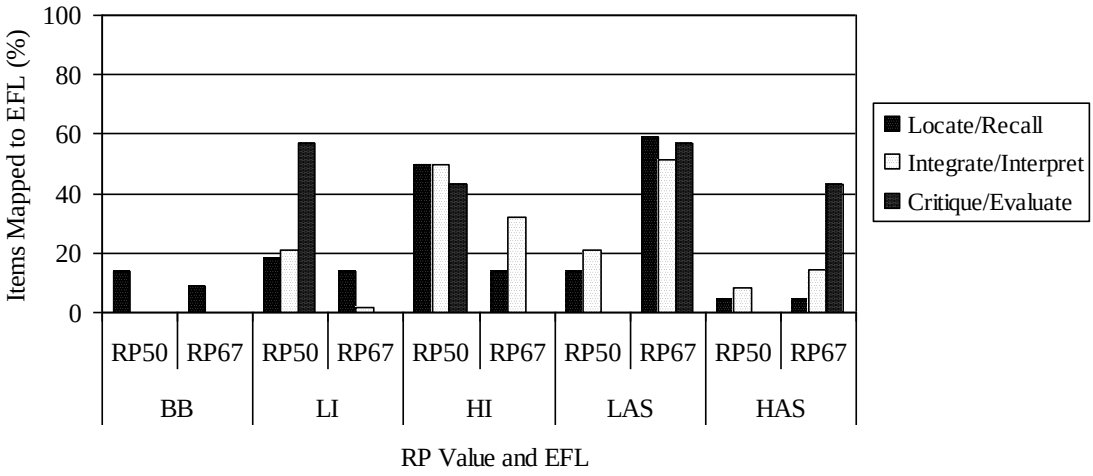


Figure 4.15. Reading Results by Cognitive Skill for High Intermediate EFL

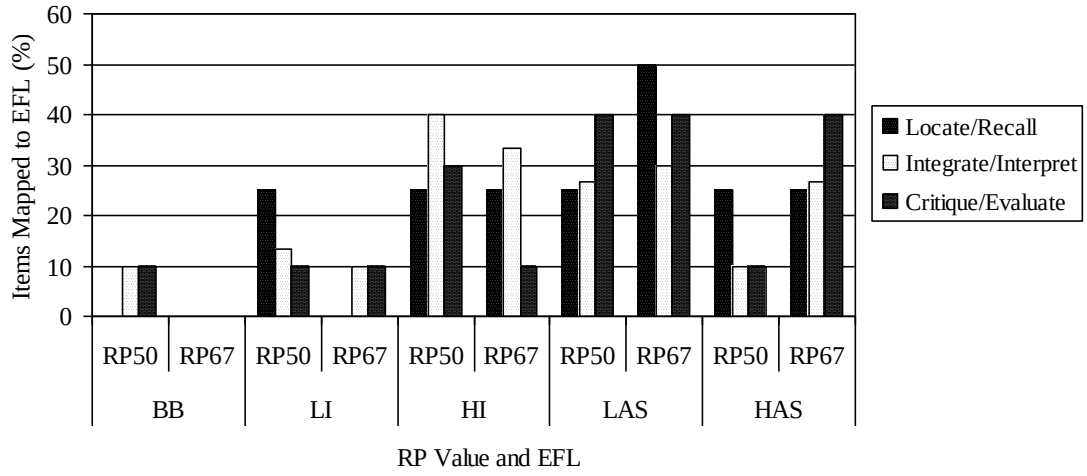


Figure 4.16. Reading Results by Cognitive Skill for Low Adult Secondary EFL

CHAPTER 5

DISCUSSION

5.1 Overview

This study was designed to illustrate how student responses to test items could be used to inform curriculum-assessment alignment. Item mapping methodology that utilizes item response theory was applied to Reading and Math assessments for adult basic education to illustrate the process. Results of item mapping were then compared to SMEs' classification of the items to evaluate the degree of agreement. This chapter discusses the major findings of the study and reference is made to the literature where possible. The first section discusses the impact of RP on item mapping results. This is followed by a discussion on the agreement between SMEs classification of the items and item mapping results. Differences between Math and Reading results are then discussed followed by comments on reasons for misalignment that SMEs suggested. The chapter concludes with outlining some limitations of the study and directions for future use.

5.2 Impact of RP Value on Item Mapping and Alignment

Results of this study show that for both Math and Reading assessments, more items tended to mapped to the lower EFLs (that is High Intermediate or lower) at RP50 while more items were mapped to higher EFLs (Low Adult Secondary or higher) at RP67. These results were expected because as pointed out earlier, most of the items used in this study had *c*-parameter values that were less than 0.35. As such, the theta value at which students have a 50% chance of providing a correct response to an item (that is RP50) will always be less than the *b*-value. The only exception to this is when the c-parameter is equal to zero. On the other hand, the theta value at which students have a 67% chance of providing a correct response to an item will always be higher than the b-

value. However, the assumption being made here is that the SMEs took the difficulty and discrimination of the item into consideration in classifying the items. The other assumption is that the SMEs estimation of the difficulty of the items for a particular group of learners was accurate. These assumptions are discussed in the next section.

Results also show that in general, greater alignment between SMEs and item mapping results was obtained at RP50 compared to RP67 for both Math and Reading. These results are similar to results obtained by Kolstad et al. (1998). In their study aimed at evaluating the impact of RP value on selection of exemplar items that could be used to describe what students at a particular proficiency level could do, the authors found the greatest agreement between the percentage of items mapped along the proficiency scale and percentage of scores for examinees along the proficiency scale at RP50.

5.3 Agreement between SMEs Classifications and Item Mapping

This study uses the degree of agreement between SMEs' classifications of the items and item mapping results as a measure of alignment. The item mapping methodology employed in this study locates the items on a proficiency scale such that easy items are located on the lower end and harder items are located on the higher end of the proficiency scale. As such, difficulty of the item plays the major role in determining where on the proficiency scale the item will be positioned. Comparing agreement between SMEs classifications of the items and location of the items on the proficiency scale assumes that some common parameter was used in the two classifications. It is hoped that SMEs consider not only the match between the item content and the level of the curriculum at which the content is taught, but also the relative difficulty of the item. As such, trustworthiness of SME's ratings of the items for the intended group hinges upon their ability to accurately judge or estimate difficulty of the item for the target

group. Research on teachers' ability to estimate item difficulty and other properties of items (e.g., discrimination) yields mixed results. Most of the research comes from standard setting realms that rely on SMEs to make judgments about the difficulty of an item for a specified group of examinees. For example, the Angoff standard setting method involves SMEs estimating the probability that a minimally competent examinee would provide a correct response to a particular item.

Impara and Plake (1998) used a survey to assess teachers' ability to estimate item difficulty. They found that teachers underestimated the performance of minimally competent students but overestimated the performance of the total group. In other words, their estimates of item difficulty were lower than the actual difficulty for the minimally competent students and higher than actual difficulty for the whole group. Similarly, Shepard (1994) found that trained panelists overestimated examinee performance on easy items but underestimated their performance on hard items.

In another study, Plake, Impara and Irwin (2000) employed about 30 well trained SMEs to estimate item proportion correct *(p-value*) for the minimally competent as well as the whole group of students. Plake et al. observed greater consistency among the panelists in estimating student performance on the hardest and moderately difficult items but the consistence was less for the easiest items. However, the observed differences were very small. The authors also observed high inter-rater and high intra-rater reliability across years and within panelists respectively. Ryan (1968) used 59 math teachers to judge item difficulty, discrimination, and relevance of 25 multiple-choice items. Results of the study indicated that teachers made relatively accurate judgments about item difficulty and discrimination especially when the content of the items was familiar to students. Lastly, Plake and Impara (2001) found substantial agreement between panelists'

item difficulty estimates for the minimally competent examinees and actual performance of examinees whose scores were within one standard deviation of the mean. In general, research shows that teachers are generally accurate in estimating item difficulty for a particular group of examinees. However, their ability to make accurate judgments about item difficulty seems to depend on other factors such as overall proficiency of the target group for which the estimates are to be made, the difficulty of the item, and the quality of training the teachers went through.

Results of the present study indicate that significant differences between SMEs and item mapping classifications of the items were observed in both Math and Reading. In general, more variance in SMEs' classification was explained at RP50 than at RP67. This observation may mean that the SMEs considered a typical student at a specified EFL as one who has at least a 50% chance of giving a correct response to an item and they used that to classify the items to EFLs. It was also observed that none of the Math items intended for the Beginning Basic EFL mapped to High Adult Secondary at RP50 and no items intended for High Intermediate EFL or higher mapped to Beginning Basic EFL. Similarly, few Reading items intended for Beginning Basic mapped 2 or 3 EFLs higher than intended and few items intended for Low Intermediate and higher EFLs mapped to Beginning Basic. This finding implies that there was some agreement between SMEs estimation of item difficulty and learners' actual performance on the items. This observation may also provide evidence on the overall ability of the SMEs to judge the difficulty of the items.

Based on these results, it appears reasonable to state that the SMEs were relatively accurate in their estimates of difficulty of the items. These results closely match results obtained by Zwick et al. (2001). In a study designed to investigate alternative item

mapping methods for the NAEP, Zwick, et.al (2001) asked SMEs to list the five easiest and five hardest items from a test without ordering the items by difficulty within each set. The authors found that the SMEs difficulty rankings matched very closely to student's performance. Specifically, a Spearman correlation between the SMEs rankings and the proportion of 8^{th} graders answering an item correctly was 0.65). Based on this correlation, Zwick et al. (2001) concluded that the SMEs "rankings were substantially in line with the actual difficulty of the items" (p. 22). Similar conclusions could be drawn about the Math results obtained in the current study. Spearman correlations between SMEs' and item mapping results were about 0.7 for both RP50 and RP67. These values are slightly higher than those obtained in the Zwick study mentioned above. On the other hand, Spearman correlation obtained for Reading was significantly low (0.48).

This study also found that at RP50 greater exact agreement between SMEs and item mapping results for both Math and Reading was obtained at the lower EFLs (that is, High Intermediate EFL or lower) while the least was obtained at higher EFLs. Considering RP67, greater agreement between the two classifications was obtained at the higher EFLs compared to lower EFLs. These results imply that most items intended for lower EFLs mapped to low EFLs while those intended for higher EFLs did map to high EFLs. This finding also provides some evidence that the SMEs made reasonably accurate judgments about items intended for lower EFLs and those intended for higher EFLs. It also provided some evidence that item difficulty was one of the item characteristics that the SMEs used to classify the items and that SMEs' estimates of difficulty closely matched actual difficulty.

The Math and Reading data were also analyzed based on the content strand that each item was intended to assess. The goal was to find out if there was any relationship

between SMEs classification of the items into EFLs and results from item mapping. Results showed that both exact as well as adjacent agreement level in Math were higher for RP50 than RP67. In other words, there was greater alignment between SMEs classification of the items and item mapping results at RP50 compared to RP67. Second, results showed no clear or definite pattern in degree of consistency of alignment across content strands and EFLs. This means that there was no particular content strand that showed consistently high agreement between SMEs and item mapping classification across the EFLs. However, it was observed that generally there was less agreement between SMEs and item mapping classifications for the Patterns, Functions, and Algebra content area. These findings may imply that the SMEs were somewhat consistent in their judgments about the items, that is, their judgments were not necessarily influenced by the content strand of the item. The SMEs' judgments about the classification into EFL of the items intended for this content strand seem to be less consistent.

Reading results by content strand were slightly different from the results obtained in Math. It was observed that higher exact agreement between SMEs and item mapping classifications were obtained for the Comprehension content strand at all EFLs except at the Low Adult Secondary EFL where Vocabulary Meaning showed the highest agreement. It could be said therefore that there was greater correspondence between SMEs judgment as regards the EFL for which Comprehension items were intended and the actual performance of the learners on the items. Conversely, such an agreement was much less for the Vocabulary Meaning content strand.

Results suggest that SMEs somehow overestimated the performance of learners on Vocabulary Meaning items at the High Intermediate EFL and lower and somewhat underestimated learner's performance on items intended for Low Adult Secondary EFL.

Evidence of this is drawn from the observation that large proportions of Vocabulary Meaning items intended for Beginning Basic for example, mapped to the next higher EFL. On the other hand, large proportions of Vocabulary Meaning intended for Low Adult Secondary mapped to High Intermediate EFL meaning Low Adult Secondary EFL learner's performance on these items was much higher than the SMEs had anticipated. The trend observed in Vocabulary Meaning items is similar to observations that Shepard (1994) made where well trained standard setting panelists overestimated examinee performance on easier items but underestimated their performance on hard items.

Considering Math results stratified by cognitive skill area, the study found that greater alignment between SMEs and item mapping classifications were observed at the Beginning Basic EFL. On the other hand, the alignment was somehow low for Evaluation cognitive skill at the lower EFLs but high at the Low Adult Secondary EFL. This observation may imply that the cognitive skill needed to answer an item as well as learner sub group had an impact on SMEs' ability to estimate its difficulty and hence the appropriate EFL. The results described above show that SMEs find it harder to estimate the performance of learners at the lower EFLs on items that require more abstract reasoning but were more accurate at estimating difficulty of such items for higher EFLs.

Item mapping results for Reading stratified by cognitive skill area show that at RP50 exact agreement was highest for the Locate/Recall cognitive skill area at the Beginning Basic, Low Intermediate, and High Intermediate EFLs. It was also observed that at the Low Adult Secondary EFL, greater congruence between actual item difficulty and SMEs estimated difficulty was obtained for the Critique/Evaluate cognitive area and the lowest was obtained at the Locate/Recall skill area. These results imply that SMEs' estimated difficulty of the items assessing Locate/Recall skills at the High Intermediate

and lower EFLs was similar to the performance of examinees on the items. On the other hand, SMEs' estimated difficulty of items assessing Critique/Evaluate items at the EFLs mentioned above were less similar to actual item difficulty as indicated by learners performance. Greater congruence between SMEs estimates and actual item difficulty of Locate/Recall items and less congruence observed for Critique/Evaluate items implies that the SMEs were more accurate in estimating the difficulty of items requiring high level cognitive skills at the highest EFL and the items requiring low level cognitive skills at the other EFLs. Similar observations were made for the MAPT for Math.

5.4 Comparison between Math and Reading Results

Comparisons between Math and Reading results reveal some interesting trends. First, it was observed that the amount of variance in SMEs classifications of the items accounted for by item mapping results at RP50 was much higher in Math than in Reading (48% vs 23%). Second, it was observed that overall exact agreement between SMEs and item mapping classification of the items was much higher in Reading (45.3%) than in Math (28.1%). Similarly, overall adjacent agreement was also higher in Reading (87.5%) than in Math (72.5%). Similar trends were observed at the individual EFLs. For example, the highest exact agreement of 34% was observed at the Beginning Basic EFL for Math while in Reading, the agreement was 49.5% observed at the High Intermediate EFL. In addition, adjacent agreement was highest at the Low Adult Secondary (84.8%) in Math while in Reading the agreement was highest at the High Intermediate EFL (90.2%). One difference between the results for the two subjects was that while the highest exact and adjacent agreement at RP50 were obtained at the High Intermediate EFL for Reading the agreement was highest at the Beginning Basic for Math. The lowest agreement was obtained at the Low Adult Secondary and Beginning Basic EFL for Reading and Math

respectively. Similar trends were observed at RP67 where both overall exact and adjacent agreement levels were generally higher in Reading than in Math.

5.5 Reasons for Misalignment

This study found several factors that could lead to items not performing as intended. Items could become easier or harder for the intended group depending on these factors. It was observed that the level of cognitive thinking that the item requires does alter item difficulty. In general items demanding higher levels of thinking were perceived to be more difficult. Analysis of the items that teachers identified as cognitively more demanding showed that they were those that the item writers classified as measuring evaluation and synthesis skills.

Teachers also identified some characteristics of the items irrelevant to the construct being assessed that could affect examinee performance. For example, difficult vocabulary, use of long sentences and excess verbiage were mentioned as some of the issues contributing to misalignment. It was interesting to note that the items reviewed in this study had content that was taught to the learners. In other words, lack of student exposure to content was not a factor that contributed to examinee low performance. It was the level of cognitive thinking the content in the item demanded that mattered most.

Test developers could improve alignment between intended and actual item difficulty by ensuring that the language in the item matches the language level of the students. This does not only improve the clarity of the item and student understanding but also eliminates construct irrelevant variance that could interfere with student performance. Another strategy would be to match the cognitive demands of the item to the cognitive capability of examinees. Matching the cognitive demand of the item to that of the students reduces the frustration and stress that might affect student performance in

an item. Alignment can also be improved by ensuring that the items are free from error. Items should be stated in simple language, and the accompanying visuals should be well drawn and well labeled where appropriate. It is also important to ensure that the distractors are plausible, that is, they cannot be easily eliminated by less knowledgeable examinees or they do not offer clues to the correct response.

5.6 Implications of Results

This study shows that utility of alignment study results could be greatly enhanced if students actual performance on the assessment can be taken into consideration. This would provide information on the strengths and weaknesses of the students and also inform teachers on which areas of the curriculum need extra emphasis. In general, results of the study indicate that SMEs are fairly accurate in their judgment of item difficulty. This study also brings to light important issues to test developers and item writers. To a test developer, results of this study offers some helpful hints that could inform training of item writers. The research brings out issues such as language level, cognitive demand and plausibility of distractors as some factors that test developers could emphasize in item writing sessions.

5.7 Limitations and Directions for Future Studies

One limitation is that the study employs a standard for evaluating alignment that is not error free. The item mapping results are compared to SMEs classification of the items in EFLs. Using SMEs classification as the criteria for judging alignment assumes that classifications were made with as little error as possible, an assumption that may not be correct. Second, the reasons for misalignment were sought for Math only. As such, it is unknown if the teachers could bring about similar issues as reasons for misalignment in other subjects. This study does not provide actual evidence of teacher instructional

practices. Instructional practices can only be inferred from the observation that lack of instruction was not cited as one of the reasons for misalignment. As such, there is no link between results of the study and teacher practices. Future studies could therefore be focused on exploring or identifying ways of assessing how much error is inherent in the SMEs classifications of the items to make accurate conclusions about alignment. Research efforts could also be directed towards incorporating teacher's instructional practices to inform alignment. Applying item mapping to subjects other than math and reading is also an idea worth pursuing.

APPENDIX A

ITEM REVIEW SHEET

Name: _____

Item ID: _____

Level written to: _____

Level mapped to: _____

Suggested reasons for misalignment

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

APPENDIX B

MAPT FOR MATH ITEM MAPPING STUDY

Survey

Dear Panelist: Thank you for your participation in the MAPT for Math content validity study. Please take a moment to give us your impression about the activities undertaken today. Your responses to this questionnaire will be kept confidential.

1. Please indicate your level of agreement or disagreement with each of the following statements about **item difficulty mapping** task. Please check the most appropriate response on the rating scale provided.

Statement	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
a. The practice exercise helped me understand the item difficulty mapping task.					
b. I feel my opinions were ignored during discussion.					
c. I clearly understood our task of providing reasons why items were more difficult or easier than expected.					
d. I had adequate time to review and comment on the item difficulty.					
e. I needed more training to confidently complete the item difficulty mapping task.					

2. What factors did you consider in reviewing the difficulty mapping of the items?

_____

_____

_____

_____

_____

_____

_____

_____

3. What suggestions do you have that could help us improve our efforts to write MAPT for Math items at the appropriate levels of difficulty?

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Thank you.

BIBILIOGRAPHY

Achieve, Inc. (2001). *Measuring Up: A report on education standards and assessments for Massachusetts.* Retrieved October 20, 2009, from http://www.achieve.org/files/MassachusettsBenchmarking10-2001.pdf

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

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

Ananda, S. (2003b). Achieving alignment. *Leadership*, *33*(1), 18-21.

Beaton A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191-204.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning Tests with content Standards: Methods and Issues. *Educational Measurement, Issues and Practice*, 2003 (22), 21-29.

Blank, R. K., Porter, A., & Smithson, S. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science*. Report from SECP(National Science Foundation REC 98-03080). Washington, DC: Council of Chief State Officers.

Buckendahl, C. W., Impara, J. C., Plake, B. S., & Haack, K. (2001). *Evaluating the alignment of selected nationally norm referenced achievement tests to Nebraska's 4th, 8th, and high school reading/writing and mathematics content standards.* Lincoln, NE; Nebraska Department of Education.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Earlbaum.

Crocker, L. &Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Belmont: Wadsworth Group.

Gall, M. D., Borg, W. R. & Gall, J. P. (1996). *Educational Research: An Introduction.* (6th ed). New York: Longman Publishers.

Gomez, P. G., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, *24*(3), 417-444.

Hambleton, R. K. (1997). Enhancing the validity of NAEP achievement level score reporting. *Proceedings of achievement levels workshop*. National Governing Board, Washington, DC.

Hambleton, R. K. & Sireci, S. G (2008). Development and Validation of Enhanced SAT Score Scales Using Item Mapping and Performance Category Descriptors.

Haynes, S. N., Richard, D. C. S, & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247.

Hendrickson, A. (2007). An NCME Instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26* (2), 44 – 52.

Herman, J., Webb, N., & Zuniga, S. (2005). *Measurement issues in alignment of standards and assessment: A case study*. (CSE Report 653). Los Angeles; University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20.

Impara, J. C. & Plake, B. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69 – 81.

Kaira, L. T. & Sireci, S. G. (2007). *Evaluating the content validity of a multistage adaptive test*. Center for Educational Assessment Research Report No. 656. Amherst, Massachusetts: School of Education, University of Massachusetts Amherst.

Kirsch. I., Jungebult, A., Jenkins, L. & Kolstad, A. (1993). *Adult literacy in America: A first look at the findings of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics; US Department of Education.

Kolen, M. (2001). Linking assessments effectively: Purpose and design. *Educational Measurement: Issues and Practice, 20*(1), 5 – 9.

Kolstad, A. (1996, April). *The response probability convention embedded in reporting prose literacy levels from 1992 National Adult Literacy Survey.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Washington, DC: American Institutes for Research.

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A. & Despriet, L. (2000). *State Standards and State Assessment Systems: A guide to alignment.* Washington, DC; Council of Chief State Officers.

Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education, 11,* 23-47.

Martone, A. & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction. *Review of Educational Research, 79*(4), 1332 - 1361.

Martone, D., Sireci, S. G., & Delton, J. (2006). *Methods for the alignment between state curriculum frameworks and state assessments: A literature review*. Center for Educational Assessment Research Report No 603. Amherst, MA: University of Massachusetts, School of Education.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed), Setting performance standards: Concepts, methods and perspectives (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates.

National Research Council. (2005). *Measuring literacy: Performance levels for adults.* Committee on Performance Levels for Adult Literacy, R. M. Hauser, C. F. Edley, Jr., J. A. Koenig, and S. W. Elliott. Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Philips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP scales.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Plake, B. S, Impara, J. C., & Irwin, P. M. (2000). Consistency of Angoff based predictions of item performance: Evidence of technical quality of results from the Angoff standards setting method. *Journal of Educational Measurement*, *37*(4), 347 – 355.

Plake, B. S. & Impara, J. C. (2001). Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting. *Educational Assessment*, *7*(2), 87 – 97).

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

Reckase, M. D. (2006a). A conceptual framework for a psychometric theory of standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25*(2), 4- 18.

Reckase, M. D. (2006b). Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. *Educational Measurement: Issues and Practice, 25*(3), 14- 17.

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 Angels, CA: National Center for Research on Education, Standards, and Student Testing.

Ryan, J. J. (1968). Teacher judgment of test item properties. *Journal of Educational Measurement*, *5*(4), 301 – 306.

Ryan, J. M. (2006). Practices, issues, trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 677-710). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Shephard, L. A. (1994). Implications for standard setting of the NAE evaluation of NAEP achievement levels. Paper presented at the Joint Conference on Standard Seting for Large Scale Assessment. National Assessment Governing Board, National Center for Educational Statistics, Washington, DC.

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

Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., Shea, C. L., Han, K. T., Deng, N., Delton, J., & Hambleton, R. K. (2008). *Massachusetts Adult Proficiency Tests technical manual: Version 2*. Center for Educational Assessment Research Report No. 677. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Schulz, E. M. (2006). Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practice, 25*(3), 4- 13.

Thorndike, R. M. (1997). *Measurement and Evaluation in Psychology and Education* (6th edn). New Jersey: Merrill.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273 – 286.

Tindal, G. (2005). *Alignment of Alternate Assessments using the Webb System*. Washington, DC; Council of Chief State Officers.

U.S. General Accounting Office (1993). Educational achievement standards: NAGB's approach yields misleading interpretations (Rep. No. GAO-PEMD-93-12). Washington, DC: Author.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement*, *40*(3); 231-253.

Webb, N. L. (1997). *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education*. Research Monograph No. 6). Washington DC: Council of Chief State Officers.

Webb, N. L., M., Ely, R., Cormier, M. &Vesperman, B. (2005). *The WEB Alignment Tool: Development, Refinement, and Dissemination*. Washington, DC; Council of Chief State Officers.

Webb, N. L (2006). *Alignment Analysis of Mathematics Standards and Assessments. Wisconsin, Grades 3-8 and 10*. Retrieved October 20, 2009, from http://www.dpi.state.wi.us/oea/pdf/mathsummary06.pdf

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International, Inc.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15-25.