Content Validity of the New York State Spanish Comprehensive Regents Examination

Lisa Mars

Follow this and additional works at: https://scholarship.shu.edu/dissertations

Part of the Curriculum and Instruction Commons, and the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
https://scholarship.shu.edu/dissertations/2671
CONTENT VALIDITY OF THE NEW YORK STATE SPANISH COMPREHENSIVE REGENTS EXAMINATION

Lisa Mars

Dissertation Committee

Christopher Tienken, Ed.D., Mentor
Reader: Barbara Strobert, Ed.D.
Thomas Tramglini, Ed.D.
Dennis Canterbury, Ph.D.

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Education

Seton Hall University
2013
SETON HALL UNIVERSITY
COLLEGE OF EDUCATION AND HUMAN SERVICES
OFFICE OF GRADUATE STUDIES

APPROVAL FOR SUCCESSFUL DEFENSE

Doctoral Candidate, Lisa Mars, has successfully defended and made the required modifications to the text of the doctoral dissertation for the Ed.D. during this Summer Semester 2013.

DISSERTATION COMMITTEE
(please sign and date beside your name)

Mentor:
Dr. Christopher Tienken

Committee Member:
Dr. Barbara Strobert

Committee Member:
Dr. Thomas Tramglini

Committee Member:
Dr. Dennis Canterbury

The mentor and any other committee members who wish to review revisions will sign and date this document only when revisions have been completed. Please return this form to the Office of Graduate Studies, where it will be placed in the candidate’s file and submit a copy with your final dissertation to be bound as page number two.
Content Validity of the New York State Spanish Comprehensive Regents Examination: Alignment to New York State Modern Languages Checkpoint B Syllabus and New York State Modern Languages Standards 2006-2011

The purpose of this alignment study of the years 2006-2011 of Part 2 of the Spanish Languages Other Than English Comprehensive Regents Examination was to describe (a) the categorical concurrence of the New York State Comprehensive Spanish Regents Examination 2006-2011, (b) the depth of knowledge consistency as defined by Webb of the items of the New York State Comprehensive Spanish Regents Examination 2006-2011, and (c) the range of knowledge correspondence of the New York State Comprehensive Spanish Regents Examination 2006-2011, defined as the number of content objectives of the New York State Checkpoint B syllabus assessed by items of the New York State Comprehensive Spanish Regents Examination.

The items were coded by a panel of raters, and it was determined that there was no consistency over time with regard to the categorical concurrence, depth-of-knowledge consistency, range-of-knowledge representation, and cut score from administration to administration of the assessment. In fact, the June 2008 administration is the only administration to have full alignment as defined by Webb for 2006-2011, the period under study.
ACKNOWLEDGMENTS

I certainly want to thank several people who accompanied me on this odyssey--my family for their love and support; Dr. Dennis Canterbury, whose daily text messages encouraged me to write; my dear friend Ines, whose focus and kind words motivated me when I was tired; my reader, Dr. Barbara Strobert, for her suggestions and feedback; and of course, Dr. Christopher Tienken, my mentor, for challenging my thinking and helping me to become a better writer and researcher.

And to all my friends and colleagues at Seton Hall University, let’s make this count in providing all students with a better opportunity to learn and to become productive citizens.
DEDICATION

This dissertation is dedicated to my family. Without your support, this journey would not have been possible. Thank you for your thoughtfulness and understanding over the past twenty-four months.

To my late grandmother--thank you for believing in me and giving me the courage to believe in myself.
TABLE OF CONTENTS

ABSTRACT ........................................................................................................................................ii

ACKNOWLEDGMENTS .......................................................................................................................iii

DEDICATION ......................................................................................................................................iv

LIST OF TABLES .................................................................................................................................ix

LIST OF FIGURES ..............................................................................................................................x

CHAPTER 1. INTRODUCTION ...............................................................................................................1

  TYPES OF VALIDITY ...........................................................................................................................3

  ALIGNMENT AND CONTENT VALIDITY .................................................................................................4

BACKGROUND OF THE PROBLEM ........................................................................................................5

  THE NEW YORK STATE LANGUAGES OTHER THAN ENGLISH REGENTS EXAMINATION.............7

  CUT SCORE OF THE NYS SPANISH COMPREHENSIVE REGENTS EXAMINATION .......................9

STATEMENT OF THE PROBLEM ..........................................................................................................10

THE PURPOSE OF THE STUDY ...........................................................................................................11

RESEARCH QUESTIONS .......................................................................................................................12

SIGNIFICANCE OF THE STUDY ...........................................................................................................13

CONCEPTUAL FRAMEWORK: EXPLORING COGNITIVE DEMAND IN ASSESSMENT ......................15

ALIGNMENT WITHIN THE CONCEPTUAL FRAMEWORK OF WEBB .................................................16

THE RELATIONSHIP OF ALIGNMENT TO CONTENT VALIDITY .......................................................17

OVERVIEW OF RESEARCH DESIGN AND METHODS .......................................................................17

LIMITATIONS OF THE STUDY ............................................................................................................19

DELIMITATIONS OF THE STUDY ........................................................................................................20

ORGANIZATION OF THE STUDY .........................................................................................................23

SUMMARY ...........................................................................................................................................23
CHAPTER 2. REVIEW OF THE LITERATURE ................................................................. 24
LANTUGUES OTHER THAN ENGLISH REGENTS EXAMINATION ...................................................... 24
TEST QUESTION EXAMPLES ................................................................................................. 26
LITERATURE SEARCH PROCEDURES .................................................................................. 27
CRITERIA FOR INCLUSION AND EXCLUSION ........................................................................ 28
CONTENT STANDARDS ........................................................................................................ 28
CONSTRUCT RELEVANT QUANTITATIVE PERFORMANCE AND STANDARDS .............................. 30
CONSEQUENTIAL VALIDITY OF STATEWIDE STANDARDIZED TESTING ................................. 34
TESTS AS POLICY INSTRUMENTS ......................................................................................... 35
VALUE INTERPRETATIONS OF TEST SCORES ................................................................. 36
TEST PREPARATION AND PEDAGOGY .............................................................................. 37
SYSTEMATIC VALIDITY ........................................................................................................ 40
CUT SCORES ....................................................................................................................... 42
ALIGNMENT ....................................................................................................................... 43
CONTENT VALIDITY AND ALIGNMENT ................................................................................. 46
SCALE OF AGREEMENT ...................................................................................................... 47
THEORETICAL FRAMEWORK: CONSTRUCTIVE ALIGNMENT AND ASSESSMENTS ................... 48
CONSTRUCTIVISM ............................................................................................................... 50
THE COMMUNICATIVE MODEL ......................................................................................... 52
CURRICULUM ALIGNMENT ................................................................................................. 54
ALIGNMENT AND COGNITIVE COMPLEXITY ...................................................................... 56
WEBB ALIGNMENT METHOD ............................................................................................... 60
ALIGNMENT METHODOLOGY ............................................................................................. 66
VALIDITY ............................................................................................................................. 69
CONTENT VALIDITY ............................................................................................................ 71
FACTORS THAT AFFECT CONTENT VALIDITY ....................................................................... 71
SUMMARY .......................................................................................................................... 73
CHAPTER 3. METHODOLOGY ............................................................................................ 75
DESIGN ............................................................................................................................... 75
CONTENT VALIDITY COMPREHENSIVE SPANISH REGENTS EXAMINATION

CUT SCORES ........................................................................................................................... 133
STANDARD ERROR MEASUREMENT .................................................................................... 135
ITEM BANK ........................................................................................................................ 136
CATEGORICAL CONCURRENCE ....................................................................................... 136
DEPTH-OF-KNOWLEDGE CONSISTENCY ......................................................................... 137
RANGE OF KNOWLEDGE ................................................................................................. 138
TRANSPARENCY OF TEST DATA REPORTS ...................................................................... 139
QUALITY CONTROL OF HIGH STAKES ASSESSMENTS ....................................................... 139
COMPLIANCE WITH FEDERAL LAW .................................................................................. 140
RECOMMENDATIONS FOR PRACTICE .............................................................................. 140
PROFESSIONAL DEVELOPMENT ..................................................................................... 140
POSITIVE WASHBACK ....................................................................................................... 141
RECOMMENDATION FOR FUTURE RESEARCH ................................................................. 141
CONCLUSIONS ................................................................................................................ 141
SUMMARY .......................................................................................................................... 142
REFERENCES ..................................................................................................................... 143
APPENDIX A: DEFINITION OF TERMS AND ABBREVIATIONS ........................................ 198
APPENDIX B: REQUEST AND APPROVAL OF RESEARCH ............................................... 203
APPENDIX C: APPROVAL OF DISSERTATION ORALS ...................................................... 204
APPENDIX D: CODING TRANSCRIPTS .............................................................................. 204
APPENDIX E: WEBB ALIGNMENT TOOL ......................................................................... 222
APPENDIX F: NEW YORK STATE CHECKPOINT SYLLABUS FOR MODERN LANGUAGES .... 223
APPENDIX G: NEW YORK STATE MODERN LANGUAGE STANDARDS ............................ 223
APPENDIX H: NEW YORK STATE SPANISH COMPREHENSIVE REGENTS EXAMINATIONS
2006-2011 ........................................................................................................................ 224
APPENDIX I: WORKSHOP MATERIALS FOR PANELISTS ............................................... 224
LIST OF TABLES

TABLE 1 LANGUAGE OF TEST FORMAT ................................................................. 72

TABLE 2 NYS SPANISH COMPREHENSIVE REGENTS ESSAY GRADING RUBRIC FOR STUDENT ........................................... 86

TABLE 3 PART 4 SCORE CONVERSION CHART. LISTED ON THE SCORING RUBRIC FOR GRADERS OF THE EXAMINATION ........................................................................................................... 87

TABLE 4 DATA RELIABILITY SET ........................................................................ 108

TABLE 5 DATA RELIABILITY SET FOR DEPTH OF KNOWLEDGE AND CATEGORY/CONTENT DOMAINS ......................... 108

TABLE 6 SUMMARY OF ALIGNMENT OUTCOMES ON EACH OF WEBB’S CRITERION, NYS COMPREHENSIVE SPANISH REGENTS EXAMINATION 2006-2011 ........................................................................... 114

TABLE 7 NEW YORK STATE REPORT CARD PASSING (%) ........................................ 115

TABLE 8 NEW YORK STATE COMPREHENSIVE SPANISH REGENTS EXAMINATIONS 2006-2011 .............................................. 117

TABLE 9 RANGE OF KNOWLEDGE CORRESPONDENCE PART 2 OF THE NEW YORK STATE COMPREHENSIVE SPANISH REGENTS EXAMINATIONS 2006-2011 ................................................................. 118

TABLE 10 THE NEW YORK STATE COMPREHENSIVE SPANISH REGENTS EXAMINATIONS 2006-2011 - CATEGORICAL CONCURRENCE ........................................................................................................ 119

TABLE 11 ALIGNMENT OF PART 2 OF THE NEW YORK STATE COMPREHENSIVE SPANISH REGENTS EXAMINATIONS 2006-2011 ........................................................................................................ 122

TABLE 12 NEW YORK STATE COMPREHENSIVE SPANISH REGENTS EXAMINATION, CONTENT VALIDITY 2006-2011 .................................................................................................................. 123

TABLE 13 NYS SPANISH COMPREHENSIVE REGENTS EXAMINATION, COGNITIVE RIGOR 2006-2011 ......................... 126
LIST OF FIGURES

FIGURE 1. BIGGS STRUCTURE OF OBSERVED LEARNING OUTCOMES (SOLO TAXONOMY) ......................... 49

FIGURE 2. CHARACTERISTICS OF ALIGNMENT .................................................................................. 54

FIGURE 3. BLOOM’S TAXONOMY WITH TASK-ORIENTED QUESTION CONSTRUCTION ............... 57

FIGURE 4. WEBB’S DEPTH OF KNOWLEDGE ................................................................................ 59

FIGURE 5. DEPTH OF KNOWLEDGE CONSISTENCY BETWEEN STANDARDS AND ASSESSMENT: MICHIGAN MEAP MATHEMATICS. ................................................................. 62

FIGURE 6. RANGE OF KNOWLEDGE CORRESPONDENCE: GOLDEN STATE EXAM IN MATHEMATICS .......... 63

FIGURE 7. NYS SPANISH REGENTS EXAMINATION JAN. 25, 2006 QUESTION 22 ................................. 97
Chapter 1

INTRODUCTION

Testing has a two-fold function--the evaluation of student achievement and the delivery of school curricula in accordance with stated objectives. Results from standardized state tests are therefore very important to help schools fine-tune their curriculum to ensure that both sides, the students and education personnel, achieve the highest possible level of efficiency in terms of student achievement and the delivery of curricula. The interdependence of student achievement and curricula delivery is undeniable (Tanner & Tanner, 1980) and state-mandated assessment is at the center of evaluating student achievement in the current policy environment of the Race to the Top and the Common Core State Standards. Policymakers use the results of standardized assessments to rate schools, make judgments about the quality of instruction, and as a basis for funding decisions.

The dilemma faced by the education bureaucrats in the United States, however, is that they are unable to formulate detailed policies on proficiency and mastery of content, promotion of students, and articulation of programs due to the unreliability of the information they receive from large-scale tests (Baker, Linn, Herman, & Koretz, 2002; Popham, 2002; Koretz, 1988). A specific problem with the results from state standardized tests of academic skills and knowledge is their imprecision of measurement and the appropriate reporting of the imprecision (American Education Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Joint Committee on Testing Practices [JCTP], 2004; Koretz, 2008; Tienken, 2011).
Although the technical limitations and flaws of large-scale testing are known, understanding the nature of the problem requires further scientific investigation. The enhancement of knowledge of the problem requires empirical explorations of the scope of NYS Spanish Comprehensive Regents Examination content and the pattern of correlation and alignment of learning standards and syllabus to statewide tests items. The alignment of the assessment items to the mandated syllabus is of importance in that the scores from standardized tests are meaningful to the extent that all students had the opportunity to learn the material tested (Bracey, 2006). There must exist a body of empirical evidence that establishes the connection between the numerical test score and the construct or trait that the underlying test items are supposed to measure. (Smith & Fey, 2000).

Education personnel, politicians, students, and the general public ought to know that the challenges of measuring student achievement are more complicated than the known limitations of existing large-scale tests. Stakeholders need this additional information in light of the fact that they expect the results from these tests will form the basis for the evaluation of student achievement and of teacher and school administrator effectiveness. Stakes for students, teachers, administrators, and schools are high because of the way state and city educational bureaucrats use the test results. For example, in New York City, bureaucrats use the results from the Regents Examinations to assign grades to schools, and test scores account for a portion of the principal’s overall effectiveness rating.

Measuring student achievement accurately involves the full range of the substantive material of the test and the relationship between test design and configuration of educational values or ideals to state-mandated syllabus and standards. Test scores are
valid in making judgments about students’ learning to the extent that the test items measure the learning expectations as defined by the state standards and mandated syllabus. Test scores are also valid in making judgments about students’ learning when there are enough test items for each skill to produce results reliable enough to make determinations about individual performance. The interpretations and decisions school administrators make about individual student achievement can be less than accurate when alignment of test items is not congruous to the state mandated curriculum in terms of skills, format, and cognitive level of difficulty.

**Types of Validity**

There are different aspects of validity. Generally, validity is the quality of an instrument to yield truthful inferences about the trait it measures. In the National Research Council report on high-stakes testing, validity is defined as the generalizability of a score to other measures of the construct domain “so that the score is a valid measure of the student’s knowledge of the broader construct, not just the particular sample of items on the test” (Heubert & Hauser, 1999 p. 75). This description of validity represents the consensus of experts in the field (Cronbach, 1971; Messick, 1994) and is codified by the professional standards of test use (American Educational Research Association, 1985; Joint Committee on Testing Practices, 1988). Validity is the single most important criterion in evaluating achievement testing and is important enough to make its way into laws and regulations. (Koretz, 2002; Lin, 1980). The provisions in the No Child Left Behind Act (2001) require that the assessments used by state education personnel to satisfy the accountability regulations, produce valid and reliable results (NCLB, 2001);
therefore, states have a legal obligation to ensure that the assessments are valid and reliable. Alignment studies are one such way to establish content validity.

A sub-category of content validity is domain representation. Domain representation is the degree to which the test adequately measures all the facets of the intended content domain. Consequently, content validity and alignment are inextricably linked.

Students who achieve proficiency on the *Languages Other Than English Examination* in New York State receive the Regents Diploma with Advanced Standing. In June 2010, 86,972 students took the comprehensive regents examination in Spanish. Of this number, 95% passed the examination by gaining a score of 65% or higher. Sixty five percent is the cut score for passing the examination to receive credit. What does the score mean in terms of what students know and are able to do? The test score is at best only a sample or an approximation of the underlying trait being measured, yet it is often accepted on faith. It gives the appearance of objectivity, neutrality, and fairness (Porter, 1995). Thus, without the alignment data, the test scores exist in a vacuum devoid of correlation to what students actually know and how much of the mandated curriculum they know.

**Alignment and Content Validity**

Alignment is the process of the coordination of three components of the education process: (a) curriculum, (b) instruction, and (c) assessment. Alignment is one important component of effective schooling and the foundation of standards-based reform (Elliott, Braden, & White, 2001; Webb, 1997; Webb, Horton, & O’Neal, 2002). Content validity studies provide evidence of the degree to which the test items measure the topical and
contentual knowledge as well as the cognitive complexity and demand emphasis the
items are supposed to measure. To establish content validity, there must be operational
definitions of content area domains established by the state syllabus and an examination
of the items and tasks on the assessment. Content area experts who examine the test
items, syllabus, and standards evaluate whether they capture the same content knowledge
and skills. The first place to start looking at validity is with the content of the test and the
alignment of that content to the mandated syllabus and standards. It is hard to argue that
the inferences from a test are valid if the test measures the content that is not included in
the state-mandated syllabus. Construct under representation, or failure to measure
accurately what ought to be measured, leads to systematic failure of the construct. A test
is a sample of the domain; and for accurate measurements, test makers need to sample
adequately from the domain implied in the construct. The alignment between the
assessment and the content it is meant to assess is an important piece of evidence in any
validity argument.

**Background of the Problem**

Although a number of states’ department personnel contemplate external exit
examinations, only two states, New York and North Carolina, had established systems at
the beginning of the 1990s. The oldest external exit examination in the country is the
New York State Regents Examination, which has been in continuous operation since the
1860s. In July 1864, the New York State Legislature passed an ordinance creating the
Regents Examination system. Since 1865, New York State has been using the Regents
Examination as a measure of curricular competency. The ordinance reads as follows:

Scholars presumed to have completed preliminary studies . . . to each scholar
who sustains such examination, a certificate shall entitle the person holding it to admission into the academic class in any academy subject to the visitation of the Regents, without further examination (NYSED, 1987, p. 1).

Panels of local teachers grade the NYS Regents Examination, and the scores appear on students’ official transcripts. In most cases, the score of a student is also averaged into his or her class grade to determine the final course grade. A college-bound student typically takes Regents Examinations in Mathematics and Earth Science at the end of ninth grade, Mathematics, Biology, and Global Studies at the end of tenth grade; Mathematics, Chemistry, American History, English, and Foreign Language at the end of eleventh grade; and Physics in the twelfth grade. Regents Examinations were voluntary until the late 1990s for students entering directly into the workforce and for schools that requested a waiver (NYSED, 1987).

On March 31, 1994, President Clinton signed Goals 2000: American Education Act into law. One of the foci of this Act was to encourage each state to develop challenging academic standards for students; however, school districts continued to have local control of these standards and their benchmarks for achievement. The Goals 2000 Act was replaced by the No Child Left Behind Act (NCLB, 2001), which established requirements for the standards and assessment systems of states. NCLB extended federally mandated testing to all student groups including all K-12 public and charter school students in Grades 3-8 and one grade in high school in the subjects of language arts and mathematics. The results from the mandated state assessments are disaggregated within each state and school by student demographic subgroups, including economically disadvantaged students, students with disabilities, and students with limited English
proficiency (LEP), major racial and ethnic groups, and gender. The rationale behind the use of these subgroups is the presumption that districts will no longer be able to hide behind school-wide averages the performance of groups that have historically poor achievement levels.

In response to this system of accountability required by NCLB, states implemented formal systems for holding school personnel accountable for test scores. High school exit exams are an increasingly used tool for accountability purposes. The number of states with exit exams increased to 28 states in 2010, with the addition of Oregon and Rhode Island. The percentage of all public school students enrolled in states administering exit exams has reached 74%. “Furthermore, 83% of the nation’s population of students of color, 78% of low-income students, and 84% of English language learners, were enrolled in public schools in states that administered exit exams in the 2009-2010 school year” (Dietz, 2010). These statistics seem to indicate that exit examinations disproportionately have an impact on poor, minority, and limited English proficient students. These statistics are a compelling reason for researchers, educators, and policymakers to ascertain that the test items align to the mandated curriculum standards.

The New York State Languages Other Than English Regents Examination

The New York State Board of Regents developed a new state syllabus, “Modern Languages for Communication” (NYSED, n.d.), in the mid 1980s. Its purpose was to define outcomes for foreign language education in terms of functions, situations, topics, and proficiencies. Outcomes were specified for three checkpoints: Checkpoint A (for one unit of Regents credit), Checkpoint B (for three units of Regents credit), and Checkpoint
C (for five credits of Regents credit). The syllabus was developed based on the functional communication modes—interpretive, interpersonal, and presentational—outlined in the American Council on the Teaching of Foreign Languages (ACTFL) *Proficiency Guidelines for K-12 Learners* (1986) and used by the Council of Europe as a basis for syllabus creation. Throughout the stages of its development, nonetheless, there was substantial input from foreign language educators around the state (NYSED, 1987; Rissel, 2006).

The last revision to the format of the Languages Other Than English (LOTE) Comprehensive Regents Examination took place in 2003. The revised Comprehensive Regents, a long-standing test for college-bound students, reflected the new communicative goals in foreign language teaching and learning. The first part of the test, two speaking performance tasks, is administered in a one-to-one setting with the student and the teacher. The second part of the test, administered at a later date scheduled by the New York State Department of Education, is comprised of three sections. Section 2, listening comprehension, is multiple choice in nature. The teacher reads a short passage, and students select the best answer to the question from the given answers in the question booklet. Section 3, the reading comprehension passage, is in the target language, Spanish, with the answers multiple choice in nature. Section 4, the final section, is comprised of two one-hundred-word written essays. The Comprehensive Languages Other Than English Examination is the gatekeeper examination to the attainment of the Advanced Regents diploma. The Advanced Regents diploma is the highest designation of diplomas awarded to high school graduates in New York State. In 2010, 38% of all graduates achieved this distinction; precisely, those figures are as
follows: 183,578 total graduates, of which 69,998 received an Advanced Regents diploma. The use of the Regents Examination test scores evolved over time, and presently education bureaucrats use the results to measure the extent to which individual students achieve the objectives of the LOTE NYS syllabus and to determine whether school personnel and students meet the required progress targets specified in the NYS accountability system for the No Child Left Behind Act.

**Cut Score of the NYS Spanish Comprehensive Regents Examination**

Kane (2010) posits that most high-stakes testing programs employ decision rules, in which cut scores play a major role. Traditionally, the cut scores were determined by analyzing the utilities (or consequences) associated with the adoption of different cut scores (Cronbach & Gleser, 1965), where the utilities were defined in terms of such outcomes as the productivity of workers hired or the performances of students placed into different courses. More recently, judgmental standard-setting studies have been designed to identify reasonable cut scores for educational assessments. The participants in a judgmental standard-setting study develop (or are given) a description of the level of competence needed for some purpose (e.g., minimal competence for licensure decisions, or basic, proficient, and advanced for some accountability programs), which is called the performance standard, and a corresponding point on the score scale, which is called the cut score (Kane, 1994). The performance standard, or achievement level, is a verbal specification of some level of achievement. Candidates with scores above the cut score have presumably met the performance standard and are assigned to the higher category. Candidates with scores below the cut score have presumably not met the standard and are assigned to a different, lower category. The cut score for the New York State Spanish
Comprehensive regents Examination is 65%. This means that a student must score above 65% to be considered passing the examination and meeting the standard. State Departments of Education rarely disclose to schools and the public what it means for a student to pass a test. Nor is there an explanation of what it means to be proficient or advanced or how the cut scores were established (Rotherham, 2006). The failure of State Education Departments to provide detailed information about the construct relevant quantitative performance “gives parents, policymakers, and the public only a partial understanding of educational progress and what measures like adequate yearly progress really mean. That’s because trying to interpret student performance on a test without understanding the passing score is like reading a map without a scale” (p. 3).

**Statement of the Problem**

Provisions in the No Child Left Behind Act (2001) specify that schools will provide, “challenging academic content standards in academic subjects that specify what children are expected to know and be able to do.” NCLB also specifies that assessments contain coherent and rigorous content and that schools encourage the teaching of advanced skills and challenging academic achievement standards that are aligned with the State’s academic content standards. The law describes “two levels of high achievement (proficient and advanced) that determine how well children are mastering the material in the State academic content standards and describe a third level of achievement (basic) to provide complete information about the progress of the lower-achieving children toward mastering the proficient and advanced levels of achievement” (NCLB, 2001).
The content students are expected to know is outlined in the New York State Languages Other Than English syllabus. Basic, proficient, and advanced levels of proficiency are referred to as Checkpoint A, B, and C on the New York State Modern Language syllabus. The syllabus is scaffolded so that the skill level builds with each year of study. Standards are the overarching curricular goals of an accountability system; and to facilitate the proper functioning of the accountability system, there must be some assurance that the tests that are used are aligned with the standards. Moreover, it is crucial that the content and level of difficulty of the test items reflect what is communicated by the standards. In other words, the tests must be aligned in the dimensions of topical and conceptual knowledge as well as cognitive complexity/demand emphasis. Alignment between tests and standards is in fact an explicit requirement of the Title I legislation (Hamilton & Koretz, 2002). No previous studies have been conducted to determine the degree of alignment between the items on the Comprehensive Spanish Regents Examination and the mandated New York State Standards and New York State Modern Languages Checkpoint B syllabus.

**The Purpose of the Study**

The No Child Left Behind Act (2001) requires that states demonstrate the alignment of their assessments to their content standards and syllabus. Alignment between content standards, syllabus, and standardized tests is a significant issue in test validation and pedagogy. The four alignment criteria are as follows: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation. Balance-of-knowledge representation is beyond the scope of this study.
The purposes for this alignment study of the years 2006-2011 of Part 2 of the Spanish Languages Other Than English Comprehensive Regents Examination are to describe (a) the categorical concurrence of the New York State Comprehensive Spanish Regents Examination 2006-2011 (categorical concurrence refers to the proportion of overlap between the content stated in the standards document and the items assessed by the New York State Comprehensive Spanish Regents Examination); (b) the depth-of-knowledge consistency of the New York State Comprehensive Spanish Regents Examination 2006-2011, which is the level of cognitive complexity as defined by Webb’s Depth of Knowledge of the items of the New York State Comprehensive Spanish Regents Examination; and (c) the range-of-knowledge correspondence of the New York State Comprehensive Spanish Regents Examination 2006-2011, defined as the number of content objectives of the New York State Checkpoint B syllabus assessed by items of the New York State Comprehensive Spanish Regents Examination.

**Research Questions**

The research questions for this qualitative study were as follows:

1. What is the categorical concurrence and proportion of the overlap, of Part 2 of the New York State Comprehensive Spanish Regents Examination 2006-2011?

2. What is the depth-of-knowledge consistency as defined by Webb’s Depth of Knowledge of the New York State Comprehensive Spanish Regents Examination items for the years 2006-2011?

3. What is the range-of-knowledge correspondence, defined as the number of content objectives of the New York State Checkpoint B syllabus assessed by
items in Part 2 of the New York State Comprehensive Spanish Regents Examination 2006-2011? The range of knowledge is defined as the number of content objectives appearing in the New York State Comprehensive Spanish Regents Examination 2006-2011.

**Significance of the Study**

This study is significant from four perspectives: (1) socio-political, (2) test validation, (3) pedagogical, and (4) theoretical and research. Porter (2002) posited that the content of instruction plays a primary role in determining gains in student achievement (Gamoran, Porter, Smithson, & White, 1997; McKnight et al., 1987; Rowan, 1998; Schmidt, 1983a, 1983b; Sebring, 1987; Walberg & Shanahan, 1983). The statement that students are more likely to learn the content that they are taught is the assumption upon which much of today’s standards-based education reform rests.

This alignment study is significant from a social-political perspective because there are usually significant consequences associated with standardized tests, such as those documented in Texas (Horn & Kincheloe, 2001) and elsewhere (Sunal & Wright, 2006). The Regents Examinations’ scores facilitate the conference of Advanced Regents diplomas, retention in grade, and teacher and school ratings; therefore, there is a need to understand what exactly is being measured. Taxpayers and parents have the right to know what the learning expectations are and whether students are meeting those expectations. Alignment between the syllabus and standardized tests should not be assumed. For example, using misaligned standardized test items as the basis for accountability can be potentially damaging for a whole community when a school is mislabeled as failing. Standardized tests that do not align with the content syllabus and
standards might unfairly assign grades to students that are not representative of their actual abilities. Findings from this study may help policymakers in the respective education systems realize potential issues in their current LOTE content syllabus and standardized tests and recommend guidelines on how to improve them.

This study is significant from test validation and pedagogical viewpoints. The standards-based approach to education expects that both assessment and instruction align with the content standards. Inadequate test coverage of the content domains of the standards can potentially not only undermine the validity of inferences from the test and about the extent to which students have mastered the standards but also misguide instruction. Research has shown that adequate coverage, exposure, emphasis, and quality of instruction related to the content covered in the test, or opportunity-to-learn, are significant predictors of students’ performance on the test (Stevens, 1997; Wang, 1998). Alignment studies can help identify potential directions toward which standardized tests may influence instruction. This alignment study will also assist colleges and administrators in making correct placement decisions in Spanish courses by providing an analysis of the construct representation of the assessment.

Scarce empirical literature exists on alignment studies and large-scale tests; and administrators, lacking knowledge of what students actually know, are unable to make fine-grained decisions to provide supports as students continue with their studies. Alignment studies are relatively new, and many theoretical and methodological issues remain outstanding (Herman & Webb, 2007). This qualitative study adds to the complement of alignment studies using the Webb method of alignment to determine alignment between the assessments and the construct it is intended to measure. This
study is significant to the teams of test makers in New York State who will create and design LOTE examinations in twenty-four languages beginning in 2012. The findings will serve as a blueprint to create assessments which have content validity and are balanced and aligned to the state-mandated syllabus.

**Conceptual Framework: Exploring Cognitive Demand in Assessment**

Over the past decades educators and psychologists have developed models for understanding cognitive complexity as it relates to curriculum and assessments. In 1956, Benjamin Bloom and a committee developed a classification of levels of intellectual behavior important in learning. This classification, Bloom’s Taxonomy, categorizes the levels of abstraction of questions commonly used in educational settings. Bloom’s committee identified three domains of educational activities. These domains are cognitive--mental, skills, knowledge; affective--growth in feelings or emotional areas (attitude); and psychomotor--manual or physical skills. The cognitive domain involves the development of intellectual skills and is comprised of six levels from the simplest, recite or recall as the lowest level, to increasingly more complex and more abstract mental levels of comprehension, application, analysis, synthesis, and evaluation. Since Bloom’s early work during the 1990s, a group of educational psychologists led by Lorin Anderson, a former student of Bloom, have updated the taxonomy. The levels ranged from remembering, understanding, applying, analyzing, evaluating, and creating. Then in 1997 Norman Webb developed a taxonomic system with labels that ranged from recall, basic application, strategic thinking, to extended thinking. Webb’s work is used by educators in test item development as well as in alignment studies to determine the degree of alignment between the state standards (curriculum) and the tests used by states.
for accountability purposes. There is no consensus in the literature regarding the use of Bloom’s Taxonomy Revised versus Webb’s Depth of Knowledge.

**Alignment within the Conceptual Framework of Webb**

La Marca (2001) wrote that the development and application of alignment methods came about from a desire to ensure that students’ test scores reflect their performance with respect to specific curricular measures. Webb developed a comprehensive and complex methodology to investigate the degree of alignment between assessment and standards. His method explores five different dimensions to understand the degree of alignment, content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability (Webb, 1999). However, the area of content focus is applied in alignment studies (Martone and Sireci, 2009). Webb’s content focus dimensions are comprised of six subcategories for analysis: categorical concurrence, depth of knowledge, range of knowledge, balance representation, structure of knowledge, and dispositional consonance. However, only the first four (categorical concurrence, depth of knowledge, range of knowledge, balance representation) are applied in alignment studies.

Webb (1997) outlined three methods for determining the alignment between the policy elements of curriculum, instruction, and assessment systems: (a) sequential development, (b) expert review, and (c) document analysis. Sequential development involves the creation and acceptance of one policy element, which subsequently serves as a “blueprint” for the creation of additional policy elements. For example, a state or district might develop academic standards for Languages Other Than English that provide guidance for the selection of a new performance-focused LOTE curriculum and
the development of performance-based LOTE assessments. The process of expert review involves convening a panel of content experts to review the policy elements and determine the extent of their alignment. Document analysis involves the coding and analysis of documents that represent the different policy elements. By integrating these three methods, test developers and education policymakers can increase the quality of the alignment process (Webb, 1997).

**The Relationship of Alignment to Content Validity**

Validity is the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of the tests (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 9). In NCLB, the proposed uses of tests include the evaluation of students’ current proficiency and their progress with regard to the state-defined standards or curricular frameworks. Hence, evaluation tests for such purposes involve the consideration of both the content of the test and the state-defined frameworks. Because alignment studies cover the aspect of content and curricular frameworks, alignment studies provide validity evidence of the test and the curriculum. Evaluating how well the test items represent the domains specified in the test blueprint is one way to make inferences on curriculum and address concerns of curriculum narrowing (Lynn, 2000).

**Overview of Research Design and Methods**

Alignment in the context of this study is “the degree of agreement between a state’s content standards for a specific subject area and the assessment(s) used to measure student achievement of these standards” (Bhola et al., 2003, p. 21). Bhola et al. (2003) and Rothman (2003) reviewed different alignment methods and posited that different
alignment methods produce different measures and that methods of alignment differ in complexity.

I used Webb’s Alignment Protocol for this qualitative case study on content validity of the Languages Other Than English Comprehensive Spanish Regents Examination 2006-2011. The Webb alignment methodology has been widely used to determine the degree of alignment between the standards, syllabus, and items on the assessment. The Webb (2007) alignment method was used to analyze the agreement between assessment items and the New York State content standard and Checkpoint B syllabus in terms of standard coverage, depth of knowledge in cognitive reasoning, range of objectives assessed, and balance among objectives assessed. The approach relied on using a panel of content experts using a read-behind method to examine content and cognitive process in the Languages Other Than English (LOTE) Comprehensive Spanish Regents Examination and, based on the experts’ ratings, calculated the statistics of alignment. The experts were selected from a pool of LOTE test contributors based on experience of over ten years teaching and writing test items for LOTE as well as educational qualifications of a minimum of a masters of arts in Spanish.

To address the first research question, categorical concurrence, the panel coded and matched the test items of the Languages Other Than English Comprehensive Spanish Regents Examination to the New York State Modern Language Standards.

To address the second research question, the depth-of-knowledge consistency, the panel coded and matched the items of the Languages Other Than English Comprehensive Spanish Regents Examination to the Webb’s Depth-of-Knowledge cognitive framework.
To address the third research question, the range of knowledge, the panel coded and matched the items of the Languages Other Than English Comprehensive Spanish Regents Examination to the New York State-mandated Checkpoint B syllabus.

For each of the alignment criteria, categorical concurrence, depth-of-knowledge consistency and range of knowledge, I calculated the percentage of alignment to determine the level of agreement (Webb 2002; Herman & Webb, 2007).

**Limitations of the Study**

A qualitative case study design is used to address the research question. By definition, case studies can make no claims to be typical. Findings are not generalizable in the conventional sense. I have no way of knowing empirically to what extent the findings are representative of the depth of knowledge of items and the alignment of items to the content syllabus and state standards beyond the years 2006-2011.

The findings of this study are descriptive and not explanatory in nature. This study can provide a description of the alignment and provide evidence or lack thereof of content validity from 2006-2011.

The researcher and the content experts are the primary instruments of data collection. The content experts are selected based on their knowledge of the Spanish language, knowledge of the content of the NYS syllabus, and the Spanish Regents Examination. The panel provides adequate evidence from the data generated to support the argument, and findings and to ensure that the study can be replicated.

There are several statistical indices that may be used to measure the amount of internal consistency for an exam. The most popular index (and the one reported in Testing & Evaluation’s item analysis) is referred to as Cronbach’s alpha. Cronbach’s
alpha is used to test the reliability coefficient for internal consistency of a survey or assessment tool. It estimates test-score reliability from a single test administration, using information from the relationship among test items. Cronbach’s alpha provides a measure of the extent to which the items on a test, each of which could be thought of as a mini-test, provide consistent information with regard to students’ mastery of the domain. In this way, Cronbach’s alpha is often considered a measure of item homogeneity; i.e., large alpha values indicate that the items are tapping a common domain. (Tucker, 1949). There is no data available on the Cronbach’s alpha results of the Spanish LOTE assessment. The LOTE Regents examination is administered in two parts. There is no test reliability data to measure the items in Part 1, which are similar to items in Part 2, and are tapping the same domain of the curriculum. This in turn can potentially contribute to error in score measurement.

A limitation to Webb’s Alignment Method is that the range-of-knowledge criteria do not readily allow researchers to examine how many ideas are combined under one objective. When an objective is broadly stated, then an item is considered matched, regardless of what else in the objective is not assessed. Combining skills within a single objective may result in an increased cognitive complexity as students are asked to do more with the range of skills, but it can result in a lower depth-of-knowledge conclusion. To obviate this, the researcher and coders will also match and code the test items against the subcategories of the themes of the Checkpoint B syllabus.

**Delimitations of the Study**

This present study is concerned only with alignment between the content syllabus for the years 2006-2011 and the corresponding LOTE Spanish Comprehensive Regents
Standardized Examination. The validity of content syllabus is a different issue and beyond the scope of this study. Within the domain of content validity, this study will not explore (1) the relevance of test responses to a behavioral universe or (2) the sampling adequacy of test responses.

There will be no study on test re-test reliability; i.e., the test is given again to the same group of students with no instruction to verify that the outcome is the same. Balance of representation of the Webb’s Alignment Method will not be researched.

Part 1 of the assessment instrument is comprised of two questions. I excluded Part 1 of the examination due to the fact that there is no way to know which two questions from the available list of 500 were selected by the student and the teacher to assess speaking skills. Parts 2 to 4 are analyzed, which are comprised of Questions 1 to 33. The results from this study represented Part 2, listening, reading, and writing skills; generalizations to the entire LOTE Spanish Examination should not be made based on the results.

Test scores in general are affected by the characteristics of the test tasks, test taker, strategies the test taker employs, and the inferences drawn from the results (consequential validity) (Messick, 1989). In this study, test scores and analysis of student responses will not be reviewed. Part 1 of the Spanish Comprehensive Regents Examination is comprised of speaking tasks for which there is no recording of student responses. In addition, language scores reflect the complexity of multiple influences such as prompts given to the speaker (gestures, body language, etc.) for which there is no visual recording. There was no analysis of final student scores, as the researcher cannot control for extensive test preparation, which often leads to misleading inferences or
construct validity and has negative effects (e.g., poor consequential validity) for some students. I will not review test scores due to the Lake Wobegone effect. The Lake Wobegone effect is named after a fictional town which appeared on the Garrison Keillor Radio Show in which all of the students were above average. There are no reliability coefficients to determine the effects of accommodations, test preparation, and ELL status on the test scores. In analyzing the scoring chart, there is a built-in scoring error margin. There is no database breakdown of the scores of the four skills--speaking, listening, reading, and writing--to diagnose the student performance on these functions.

There are no data available to conduct an item analysis of the student performance on the test. Reviewing each test individually through item analysis can prove to be extremely useful in improving the objective of the language tests. Item analysis deals with two factors to determine the effectiveness of an exam: the facility value (F.V.) and the discrimination index (D.I.). According to Alderson, Clapham, and Wall (1995), “F.V. measures the level of difficulty of an item, and the discrimination index measures the extent to which the results of an individual item correlate with results from the whole test” ((De Benedetti, p. 80). Analysis of test scores is beyond the scope of this study.

I used Webb’s framework of cognitive complexity or demand emphasis. Cognitive complexity involves content match and depth match. Balance-of-knowledge representation is beyond the scope of this study.

I selected the years 2006-2011 to study the issue of content validity and alignment. There are two administrations of the test, one in January and one in June. Parts 2 to 4 will be analyzed (Questions 1-33), a total of 396 items.
Organization of the Study

Chapter 1 introduces and addresses the problem and provides the framework for the study. Chapter 2 reviews the literature and theoretical framework of alignment. Chapter 3 provides an explanation of the methods used to conduct the study. Chapter 4 presents the results of the data collection, and Chapter 5 offers conclusions and recommendations for policy and practice.

Summary

This explorative qualitative study seeks to establish the alignment of the Spanish Comprehensive Regents examination to the required syllabus. It seeks to establish the extent to which the testing instrument is aligned to the New York State standards and syllabus so that correct inferences can be drawn from student test scores and other fine-grained articulation and policy decisions can be made by administrators and policy-makers.
CHAPTER 2

REVIEW OF THE LITERATURE

The purpose for this literature review was to examine the empirical literature on content validity and alignment between standardized testing instruments and the content from the syllabus they are intended to measure. Issues related to assessment and alignment, such as construct validity and the theoretical framework for alignment, are also discussed. The review of the literature helps establish a conceptual and theoretical framework for this study. This chapter begins with a historical overview of the Languages Other Than English Examination, leading to literature search procedures and criteria for inclusion and exclusion. The literature review continues with content standards, consequential validity of statewide-standardized testing, alignment, theoretical framework of alignment and alignment studies, content validity, and factors that affect content validity.

Languages Other Than English Regents Examination

State education personnel across the nation are contemplating the implementation of external exit examinations; however, the education systems of two states, New York and North Carolina, established exit examination systems at the beginning of the 1990s. The oldest external exit examination in the country is the New York State Regents Examination, which has been in continuous operation since the 1860s. Panels of local teachers grade the Regents examination, and the results from those assessments appear on students’ official transcripts. In most cases the scores from individual students are averaged into their class grades to determine the final course grade as well. A college-bound student typically takes Regents Examinations in Mathematics and Earth Science at
the end of ninth grade, Mathematics, Biology, and Global Studies at the end of tenth
grade; Mathematics, Chemistry, American History, English, and Foreign Language at the
end of eleventh grade; and Physics in the twelfth grade. Regents Examinations were
voluntary until the late 1990s for students entering directly into the workforce and for
schools that requested a waiver.

In 1993, a coalition of four national language organizations, (the American
Council on the Teaching of Foreign Languages, the American Association of Teachers of
French, the American Association of Teachers of German, and the American Association
of Teachers of Spanish and Portuguese) received funding to develop standards for foreign
language education, Grades K-12. This was the seventh and final subject area to receive
federal support to develop national standards as part of the George H. W. Bush
administration’s America 2000 education initiative, which continued under the Goals
2000: Educate America Act in the Clinton Administration. As a result of the new
content and learning standards, the Regents assessments were redesigned under the
supervision of the three divisions in the NYSED: subject matter bureaus, the curriculum
office, and the testing office. One function of the subject matter experts was to manage
the committees of classroom teachers who traveled to Albany and provide training on
writing test questions. The teachers then spent part of the summer (on a paid basis)
writing these questions. Other teachers participated in the process by “testing the test.”
They were asked by the NYSED to give randomly assigned pre-tests to their Regents
classes, prior to the Regents Examinations as part of the process to judge the difficulty of
the test items. Thus, each Regents Examination was the culmination of a long
developmental procedure. In 1994, the NYSED adopted new New York State Learning
Standards, the revised Regents Examinations, and mandated the examinations for all general education students. Accordingly, the LOTE Regents Examinations were revised and the redesigned test administered in 2003. Starting in 2011, citing lack of funding and changes to the federal policy with the introduction of Race to the Top by the Obama administration, NYSED personnel no longer produce and distribute any of the LOTE examinations. The responsibility to create, distribute, and administer the LOTE examinations now resides with the local districts, which must adhere to the same procedures previously in place for test construction and administration. The law requiring students to take a LOTE examination for the conference of the Advanced Regents diploma was not changed.

**Test Question Examples**

An example of test questions on Section 3 of the June 2009 Spanish Comprehensive Regents Examination (first paragraph) is as follows:

**Dolores Huerta: el trabajo de una vida**

Se dice que los ojos muestran mucho de la experiencia de una persona. Cuando miramos los ojos de Dolores Huerta vemos la historia de una vida llena de tristezas y de triunfos. Sus ojos inspiran fuerza y esperanza, aspectos centrales de la misión de Dolores Huerta. Aunque tiene más de 75 años, la cofundadora de la Unión de Campesinos Unidos todavía es muy activa. Ella continúa organizando a los trabajadores y defendiendo los derechos de todos.

The corresponding question is No. 16, as follows:

16. ¿Qué hace todavía Dolores Huerta a su edad?

(1) Continúa escribiendo su biografía.
(2) Continúa asistiendo a la universidad.

(3) Continúa trabajando en una oficina de salud pública.

(4) Continúa ayudando a los trabajadores agrícolas.

The correct answer is Number 4, found in the last two lines of the first paragraph; specifically, “... cofundadora de la Unión de Campesinos Unidos todavía es muy activa ...” The answers to the questions are found in sequence in the reading passage. In the question, the word todavía appears. This leads the student to that line in the passage in the first paragraph in which todavía appears and is the answer. Matching the synonyms, or closely coded words of campesinos, translated as farm workers to trabajadores agrícolas, translated as “agricultural workers,” derives the answer. The word agrícola looks like the English word agriculture and the word campesino is a word taught at the Checkpoint A. The ensuing qualitative study will analyze the alignment of the examination to the Checkpoint B syllabus and the cognitive complexity of the test as indicated by Webb’s Depth of Knowledge.

**Literature Search Procedures**

The literature review for this chapter is accessed using the search engines of ERIC, EBSCOhost, Proquest, and JSTOR as well as online and print editions of peer-reviewed educational journals such as AERA journals. Sources included books, peer-reviewed articles, journals, university dissertations, and government reports. Key search terms were reductionism, standardized testing, alignment, exit exams, validity and reliability. Spanish Comprehensive Regents Examinations were obtained from the archives of the State of New York Library. The state standards and the syllabus were obtained from the New York State Department of Education, Office of Assessment,
Policy Development, and Administration. I followed the framework for scholarly literature reviews developed by Boote and Beele (2005).

**Criteria for Inclusion and Exclusion**

The purpose of this literature review was to examine the relevant literature as it relates to the research questions. The criteria for inclusion were the following:

1. Qualitative case studies; quantitative studies related to the research questions that used correlational, quasi-experimental, or experimental designs were included where applicable
2. Articles published in peer-reviewed journals (because peer-reviewed adds a layer of academic strength and integrity)
3. Conceptual articles published in peer-reviewed journals to aid in the understanding of the scope of the problem
4. Books for seminal works and underlying theories
5. Publication dates between 1960 and 2012
6. Studies of large-scale assessments Grades 8 to 13, specifically exit examinations
7. Studies that use Bloom’s Taxonomy or Webb Alignment Method

Exclusion was limited to one criterion: Linguistic and psychological literature

**Content Standards**

Content standards have led to an increase in large scale standardized testing, as standards are part of an accountability system. Absent from the discourse on standardized testing is who is most affected by the consequences of testing every public school child in multiple subjects.
Diane Ravitch, (1995) stated that the following:

Discussion of standards tend to turn into debates about testing—such as whether tests are fair (however that word may be defined), whether tests discriminate against disadvantaged or minority students, whether test items are culturally biased by their vocabulary, whether multiple-choice tests discourage creativity, whether tests can measure what is really important, whether tests have too much influence on instruction, and whether tests should influence decisions about college admission and employment (p. 11).

However, standards are the model for what has to be achieved, the framework of learning—an example to be aimed for. Standards do not mean standardization, however; the increasing use of external and predominantly multiple-choice examinations to measure achievement of the standards is of concern to some educators.

For the purposes of this study, standards will be defined in the following way: content standards or curriculum standards give a description of what teachers are supposed to teach and what students are expected to learn. Content standards are defined by the National Education Goals Panel as “those skills which include ways of thinking, working, communicating, and reasoning, and investigating that characterizes each discipline and knowledge of the most enduring ideas, concepts, issues, dilemmas and information of the discipline” (Technical Planning Group, 1993). These enduring ideas are outlined in the state syllabus for each discipline. Content standards should be specific, clear, and measurable so that teachers can design lessons appropriate to what students should know and be able to do.
The modern languages learning standards were instituted in 1996. The comprehensive Regents Examinations “clarify that students have obtained the knowledge and skills specified by the learning standards. Very simply, what was an implicit domain has become an explicit domain that is given operation definition by the learning standards and the derivative performance indicators adopted by the Board of Regents” (DeMauro, 2002).

Standards are one of the currencies of education reform. The National Education Goals Panel, created in 1990 by the Goals 2000 legislation, was an independent executive branch agency of the federal government charged with monitoring national and state progress toward the National Education Goals. One of these goals was that each state would have “world class academic goals” that would serve to raise student achievement, and then mandate tests that closely matched these standards. Based on the European and Asian experience, advocates maintained that the annual release of school-by-school performance data would by itself create irresistible pressure on the schools to find effective curriculum materials, implement effective instructional strategies, and do whatever else was needed to raise student performance. The National Education Goals Panel was disbanded by congress when NCLB was signed into law in 2002. By this time, each state had set its own standards for student achievement in content areas and created its own tests and scoring systems.

**Construct Relevant Quantitative Performance and Standards**

The rationale for construct validation (Cronbach & Meehl, 1955) developed out of personality testing. “There is no uniquely pertinent criterion to predict ego strength, nor is there a domain of content to sample. Rather, there is a theory that sketches out the
presumed nature of the trait. If the test score is a valid manifestation of ego strength, so conceived, its relations to other variables conform to the theoretical expectations.” (Cronbach, 1971, pp. 462–463)

There is a need as in all assessments for evidence of construct validity. The items on the assessment must demonstrate evidence and theory to support the interpretations of test scores entailed by the proposed uses of the tests. These test scores or quantitative performance indicators exist in relation to the mandated standards, and I discuss the salient role of both positive and negative consequences. Cronbach and Meehl’s (1955) article on construct validity provided an expansion of validity from earlier notions of how well the test does the job it is employed to do to the notion of interpretation of test scores and consequences resulting from various uses of assessment. There must exist conceptual links between a test taker’s performance on an assessment and other variables such as “a qualitative description or quantitative score, an interpretation of the ability under assessment, and the decisions that will be made, and the consequences of the decisions that are made” (Bachman & Palmer, p. 30). Given the high stakes of many uses of assessments (schools, teacher accountability, employment decisions), the discussion of consequences is important. Kane (2006) posits, “To validate an interpretation or use of measurements is to evaluate the rationale, or argument for the proposed conclusions and decisions . . . ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified” (Kane, 2006, p. 17).

If educational personnel and policymakers add a generalization inference by taking the observed score as an indication of expected performance over a domain of
similar task performances, they create a broader and more interesting interpretation and incur an obligation to show that the scores are indeed generalizable over tasks (through appropriate reliability or generalizability studies). They can extend the interpretation further by assuming that the scores are also generalizable over occasions and/or contexts and that these extensions would introduce additional invariance assumptions that need to be examined. They can add meaning to the scores by referencing them to norms for different groups or to performance levels, benchmark performance levels (standards), or achievement levels (as in NAEP or CEFR) (Messick, 2010).

The New York State Compressive Report Card 2010 indicated a passing rate or average score of students with a score greater than 65% on the Comprehensive Languages Other Than English exams as 95% French, 97% German, 99% Hebrew, 98% Italian, 97% Latin, and 95% Spanish. The New York State Comprehensive School Report Cards from 2006, 2007, 2008, and 2009 reflect similar statistics.

In “Setting Educational Standards High Enough,” Messick (1997) argued that when high stakes are attached to tests, states have generally set low standards for these tests.

When states have required that a test must be passed to earn a high school diploma, nearly every state has pegged the high school graduation test at a ninth grade level, or less, resulting in typical passing rates of well over 90%” (p. 6).

James Popham (2002) argued that schools and even whole states could make steady gains on standardized tests without offering students intellectually challenging tasks and that test-prep activities—not authentic teaching and learning—were responsible for much of the increase in test scores which threatens, ipso facto, validity of the scores.
This explains why achievement gains on state tests are often at odds with stagnant performance on the National Assessment of Educational Progress (Cavanaugh, 2007). It explains why higher passing rates on standardized tests have had little effect on the high proportion of students who enroll in remedial college courses (Fitzhugh, 2007; Kollars, 2008). New York City’s four-year graduation rate has been rising under the Bloomberg administration, reaching 61% in 2010, the most recent year for which there are data. Also rising, however, is the percentage of students who require remediation in math and English when they get to a City University of New York college: 22.6% in 2010, up from 15.4% in 2005. (NYCDOE Office of Accountability).

Data from the NYC Department of Education Progress Report claim that in the 2010-2011 school year, 37.2% of graduates were college-ready. The cohort data showed that 24.7% are graduating college-ready based on the College Readiness Index (CRI). The College Readiness Index is defined as the percentage of students in the 2011 cohort who met the standards for passing out of remedial coursework at CUNY, by: graduating with a Regents diploma, and earning a 75 or higher on the English Regents or scoring 480 or higher on the Critical Reading SAT, and earning an 80 or higher on one math Regents and demonstrating completion of coursework in Algebra II/Trigonometry or a higher-level math subject, or scoring 480 or higher on the Math SAT. A school’s College Readiness Index accounts for 10% of a school’s grade on the progress report.

Researchers from the Manhattan Institute Center for Civic Information claim that nationally only 32% of students leave high school academically prepared for college (Greene & Foster, 2003). This percentage is even lower among Black and Hispanic students (20% and 16%, respectively). These staggering figures are especially
disconcerting because these students are more likely to need remediation in college--and are far less likely to complete a degree--than classmates who enter with higher skill levels. Note also that college remediation courses are a huge business for colleges. Nonetheless, ultimately, not having a college degree means these individuals will have a harder time finding meaningful work in today's knowledge economy. Concerns about college and career readiness and the variation of standards between states have led to the renewed call for national content standards. On June 2, 2010, The Council of Chief State School Officers (CCSSO) and the National Governors Association Center for Best Practices (NGA Center) presented the Kindergarten-12 Common Core State Standards documents. Forty-eight states, two territories, and the District of Columbia have adopted these standards. These Common Core State Standards apply to the disciplines of English Language Arts and Mathematics and are said to represent a set of expectations for student knowledge and skills that high school graduates need to master to succeed in college and in careers. New tests are in development to align with these new standards. These tests also raise concerns of equity and access.

**Consequential Validity of Statewide Standardized Testing**

Messick (1989, 1996) provided an expanded definition of construct validity by using the term consequential validity as a subcategory of construct validity. Messick (1989) posited, “Construct validity evidence provides the ‘evidential basis’ for test score interpretations. The use of a test for a specific purpose is justified by general construct validity evidence as well as its relevance and utility. The value interpretations attributed to the test construct constitute the consequential basis for test interpretation because value implications are considered an integral part of score meaning” (p. 38). Messick defines
consequential validity as “encompassing concepts ranging from the uses of tests, the impacts of testing on test takers and teachers, the examination of results by decision makers, and the potential misuse, abuse, and unintended usage of tests.” Consequential validity implies that tests have various influences both within and beyond the classroom and refers to the societal implications of testing that make up one facet of test validity.

**Tests as Policy Instruments**

In addition to the role of testing in comparing test-takers' scores to a pre-determined norm group to discriminate and determine rank, (Koretz, D. 2008), student assessments act as policy instruments and monitoring devices of different types, as there are consequences attached to students’ and schools’ performance. State education personnel currently use the data from state assessment results for a variety of purposes, including evaluating students, holding schools accountable for making annual yearly progress under No Child Left Behind, and rewarding or sanctioning individual teachers and administrators through merit pay and end-of-year ratings, which can lead to continued employment or termination.

A key concern of psychometricians is the multiple uses of a single test’s score. The American Educational Research Association’s policy statement on high-stakes testing declares, “Each separate use of a high-stakes test, for individual certification, for school evaluation, for curricular improvement, for increasing student motivation, or for other uses requires a separate evaluation of the strengths and limitations of both the testing program and the test itself” (American Educational Research Association, 2000). A report issued by the National Research Council in 1999 similarly cautioned that tests must be shown to be valid for each of the separate ways in which they are used (Heubert
& Hauser, 1999). Despite the concerns of testing specialists, the same tests are often used for a variety of policy functions for which their validity has not been established.

**Value Interpretations of Test Scores**

The assessments also have “positive consequences,” such as honors diplomas or other rewards for students who do well. Some such rewards pose issues of fairness and appropriate test use. Some of the negative consequences for students who do not do well include retention in grade or ineligibility for graduation. As a result of the positive and negative consequences outlined above, standardized tests such as the Regents Examinations, affect what content is taught, when it is taught, and the level of cognitive difficulty and format of the content. The Regents Examination also influences what is added to or deleted from the curriculum to achieve scoring benchmarks. Koretz (2005) posits that inappropriate test preparation leads to inflation of test scores due to teaching to the test and inappropriate pedagogy. Inflation gives erroneous curricular inferences based on the test scores achieved. That is, the test scores give a mistaken conclusion about what students know based on the items of the content assessed.

Because the data from these tests are collected and used for different purposes such as ranking students, evaluating teachers, monitoring curriculum, determining value added to the learner, and assigning grades to schools, it leads the public to the inference that the test score is correlated to effective instruction and learning. Furthering the assumption that the test is correlated to good instruction and learning is that when testing becomes so high-stakes that teachers’ jobs, schools rankings, and such are linked to these assessments, the classroom environment can become a test preparation factory, with instruction focused narrowly on gaming the test and commercial test preparation material.
substituted for the curricula. The decline in teaching and learning is neither an aberration nor a malfunction of the system but rather what happens when the system is working. A meta-analysis of research on test preparation showed that the effect of coaching students on items that parallel those on the test had a 0.23 to 0.45 standard deviation effect on the test taken. (Kulik, Kulik, & Bangert-Drowns, 1984). As cited by Hattie (2009), Fuchs and Fuchs (1986), in a study on Effectives of Systematic Evaluation of Exceptional Children, conducted research with an effect size of 34 and found the average effect size to be 0.28 and the percentile gain to be 11 points. Subsequent studies by Gocman (2003) with an effect size of 233 and Kim (2005a) with an effect size of 644 and 622 all produced significant results above the 0.05 level, 0.40 and 0.39, respectively (The percentile gains were 15 and 16, respectively). This means that this type of test preparation leads to higher scores on the particular measure. Teachers’ familiarity with the test can lead to consciously or unconsciously teaching material similar in format to the questions on a particular test. These two factors can boost test scores rather than the underlying student learning it is intended to measure.

**Test Preparation and Pedagogy**

Under the heading test preparation and pedagogy, the focus is on how students, and teachers are affected by the test scores and the implications for educational institutions.

The New York State Report Card (accountability and report overview 2009-2010) shows that African-American and Hispanic/Latino students did not make annual yearly progress. In New York ethnic groups tend to live in concentrated geographical areas. As a result of geographic location, schools tend have large populations of a specific ethnic
group or groups. A large-scale survey of testing practices (Madaus et al., 1992) confirmed that schools with high rates of poor and minority children were forced to engage in the highest levels of narrow and reductive instruction due to test preparation. Schools with large concentrations of such children experience the negative consequences much more than advantaged schools in that they are pressured more to raise scores and have to redirect instructional energies toward test preparation (Madaus, West, Harmon, Lomax, & Viator, 1992). With the majority of resources in historically poor and underfunded districts being directed to test preparation, this leaves no money and time for field trips and extra curricular activities. In low-performing schools the problem of inadequate quality of content has not been addressed and leaves this population more vulnerable to a curriculum of test preparation. The schools then focus on measured outcomes rather than improving teaching and learning (Darling-Hammond, 1991; Haertel, 1989; Hartel & Calfee, 1993; Haertel, Ferrara, Korpi, and Prescott, 1994; Jaeger, 1991; Koretz, Lin Dunbar, and Shepard, 1991; Madaus, 1991; Shepard 1991; Stake, 1991).

High-stakes tests have other adverse effects, one of which is the “deskilling” of teachers and directing teachers to peruse the content for test items; thereby, the test items act as curricular magnets (Popham, 2002). In test preparation programs, the burden of tests and test preparation subtracted substantially from instructional time, drilling on test preparation materials displaced topics and whole subjects, restricted pedagogy to test-like learning activities, and teachers were deskilled (Smith & Rottenberg, 1991). The “deskilling” becomes evident when there is teaching to the test versus teaching for knowledge. This act narrows the curriculum, which then becomes isolated facts and
skills. What do these isolated facts and skills enable students to do? Are the items on the test of sufficient breadth that it covers the spectrum of the expectation of learning laid out in the state syllabus? Cohen et al. (1993), Gardner (1991), and Wiske (1998) posited that students learn best when they construct their own knowledge, as tasks are made impossible when information is far removed from their connections to it. Since the teacher is using mass produced test preparation material, the curriculum is now divorced from cultural references unique to the student’s experiences in his or her community and the classroom. Overused test preparation and technical renderings of standardization drain the human content from the information conveyed and mastered. That loss is even more severe for minority students whose culture is absent from the content of standardized curricula.

Standardized tests assess cognitive skill levels easiest to measure. What is assessed becomes what is valued, which becomes what is taught (McEwen 1995a, p. 42). The tests being administered are mostly multiple choice or short answer and do not take into account the myriad of ways students learn, nor does it take into account the epistemologies of the subject. Multiple choice by its nature neglects the fact that students construct meaning and connect course content with prior knowledge. Students can only select given answers, and continual test preparation exposes students to the wrong answer three-fourths of the time. That is, a student will spend three-fourths of his or her time on information he or she is expected to forget. Regardless of the type of exercise--multiple choice questions, essays, performance tasks and such--tests represent a sample of student knowledge and should be compared to the content universe to establish validity and reliability.
A consequence of this reductionist curriculum and teaching practices is that students also have to be taught to recognize the format of the test, and the test-driven curricula become a proxy for education. This is where the effects of high-stakes testing on low-performing and minority students begin to skew the possibilities of their access to a richer education. There is so much attention to the test scores that the issue of content knowledge and skills is absent from the discourse and the rise in test scores justifies more test preparation.

**Systematic Validity**

In 2009, the U.S. Secretary of Education, Arne Duncan, announced that states with leading school reform were eligible to compete for $4.3 billion dollars in competitive grants. According to the press release for the USDOE in July 2009, the fund has four major areas in which states must show evidence of education reform:

- Adopting internationally benchmarked standards and assessments that prepare students for success in college and the workplace
- Recruiting, developing, rewarding, and retaining effective teachers and principals
- Building data systems that measure student success and inform teachers and principals how they can improve their practices
- Turning around the lowest-performing schools

New York State secured one of these competitive school reform Race to The Top grants. As a result, there was a change in the teacher evaluation policy. In 2012, NYSED made changes to teacher-related policies by tying teacher evaluations to student-
achievement data. In this manner, tests will be used as one way to fire teachers without extensive paperwork and to circumvent tenure. *Caveat emptor* when tests are referred to as accountability measures and not in terms of academics. The use of test scores for accountability stops the discourse on the quality and purpose of education since what appears on the surface to be data of students’ learning on closer examination is a fundamental realignment of power relations governing education. This policy brings to the discourse the question of systematic validity (Berry, 1994; Cohen, 1994; Frederiksen & Collins, 1989; Pierce, 1992) or test impact (Bachman & Palmer, 1996). When there is a lack of systematic validity, it means that there are factors that push measurements in the same direction. These results cannot be replicated across trials. Donald Campbell (1979) commented that “the more any quantitative social indicator is used for decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor” (p. 68).

Messick highlighted social consequences as an aspect of construct validity but argued that “If the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced . . . then the validity of the test use is not overturned” (Messick, 1989, p. 88). In this framework, adverse consequences serve mainly to suggest that some sources of invalidity (those linked to adverse consequences) are more serious than others. In contrast, Cronbach (1971, 1988) maintained that negative consequences could invalidate test use even if the consequences cannot be traced to any flaw in the test because “tests that impinge on the rights and life chances of individuals are inherently disputable” (Cronbach, 1988, p. 6). He argued that “Validators have an obligation to
review whether a practice has appropriate consequences for individuals and institutions, and especially to argue against adverse consequences” (Cronbach, 1988, p. 6).

**Cut Scores**

Rotherham (2006) states that states too rarely explain what it actually means for a student to pass a state test, to be “proficient,” or how passing scores are established. This gives parents, policymakers, and the public only a partial understanding of educational progress and what measures like “adequate yearly progress” really mean. The cut score of 65% for the New York State Comprehensive Regents Examination was determined by NYSED as part of a new policy to increase academic rigor. NYSED personnel determined that 65% must be the cut score for all Regents Examinations. The cut scores themselves and how they are set is an important discussion in standard setting, as the cut score codifies what it means for a student to pass or be proficient. In fact, the entire issue of cut scores and the process by which they are set is rarely a focus of much public or media attention at all. This leaves the public with an incomplete picture and understanding of the difficulty of passing a test. The cut score is essential to making sense of student scores, state educational progress, NCLB requirements, and various claims and counterclaims about student and school performance.

Samuel A. Livingston and Michael J. Zieky (1982) describe the judgments standard setters must make:

Any standard--absolute or relative--is based on some type of judgment. A standard is an answer to the question “How good is good enough?” and this question can only be answered by someone’s judgment. The choice of a passing score will involve judgments at some point in the process. It is important that
these judgments be (1) made by persons who are qualified to make them, (2) meaningful to the persons who are making them, and (3) made in a way that takes into account the purposes of the test. These three requirements are interrelated, and the different methods for choosing a passing score require different types of judgments and, therefore, somewhat different qualifications for the judges (p. 12).

When interpreting cut scores, it is essential to remember that they are meaningless outside the context of a specific test. Therefore, comparing cut scores between tests or states is futile because, for instance, a cut score of 33 out of 50 on one test may or may not be more demanding than a cut score of 27 out of 50 on another test. If a test has a low cut score, the media and other observers should look closely at the test. But a low cut score may not always be a bad thing. It might be an especially difficult test. In the same way, high cut scores do not guarantee rigor. It may be a very easy test. The NYS Spanish Comprehensive Regents Examination is a combination of raw scores for Parts 1, 2, and 3 and a scaled score for Part 4.

Alignment

NCLB 2001 mandated that states set standards and create assessments that are aligned with them. Today there seems to be the misguided sentiment that educators have not set the bar high enough; and as a result, students in the United States are falling behind other countries as measured by TIMMS (Trends in International Mathematics and Science Study) and PISA (Programme for International Students Assessment). The development of the CCSS (Common Core State Standards) for English Language Arts and Mathematics is one step towards nationalizing what students are supposed to know
and be able to do and is partially a response to performance on these examinations, with the intention of raising the academic standards.

Implicit in the theory of creating centralized standards is that the instruction and assessment will follow suit. For this to happen, there needs to be alignment between the curriculum, the assessment, and the standards. Current research makes a persuasive case for creating the opportunity to learn for all students (Boykin & Noguera, 2001; Hoy, 2010), as this leads to self-efficacy, which has functional value with respect to academic outcomes (Fas et al., 2010; Schultz, 1993; Williams & Williams, 2010).

*Webster’s Dictionary Online* defines alignment as “a forming in a line, to bring components into proper coordination.” Two or more system components are aligned if they are in agreement or match each other. The most common educational use of the concept of alignment refers to the match between an assessment instrument (or instruments) and a curriculum. Curriculum alignment defined as articulation, connections, and linkages also refers to the degree to which the curriculum builds across grades horizontally and supports and reinforces prior learning. Here, alignment is analogous to instructional or curricular validity of a test, (Harmon, 1991, as cited in Webb, 1997). Curriculum alignment is also defined as a comprehensive approach to teaching and learning that goes beyond any single measure of the curriculum taught or learned (Shepard, 1991b, 1993). Curriculum alignment is based in the doctrine of no-surprises; that is, children will not be taken by surprise regardless of the form of the assessment, as assessment is an integral part of the instructional program and not an add-on. (Wraga, 1999). Curriculum alignment is “the degree to which expectations [standards] and assessments are in agreement and serve in conjunction with one another
to guide the system towards students learning what they are expected to know and do” (Webb, 1999).

“Pioneered by Andrew Porter and Norman L. Webb, systematic procedures for assessing alignment have been well developed (Ananda, 2003; Bhola, Impara, & Buckendahl, 2003; Herman, N. M. Webb, & Zuniga, 2003, 2005; Olson, 2003; Porter & Smithson, 2001; Rothman et al., 2002; Webb, 1997, 2002, 2005) and now are applied in states across the country” (Webb, N. M. Herman, & Webb, N.L., 2007, p. 17). In essence, these approaches convene panels of experts to analyze assessment items against a matrix defined by an exhaustive set of topics comprising a subject area domain and by levels of cognitive demand, reflecting a range from rote memory to procedures, applications, and complex problem-solving. The matrices then become the basis for computing various indices of alignment to convey how well a test reflects intended standards. (Webb et al., 2007).

Alignment should be studied across time, using several instruments to give a consistent picture of common understanding of what students learn, to provide consistent implications for instruction, to be fair to all students, and to be based on sound principles of cognitive development. Depth-of-knowledge consistency varies on a number of dimensions, including the level of cognitive complexity of information students should be expected to know, how well they should be able to transfer this knowledge to different contexts, how well they should be able to form generalizations, and how much prerequisite knowledge they must have in order to grasp ideas. The depth of knowledge or the cognitive demands of what students are expected to be able to do is related to the number and strength of the connections within and between mental networks (Hiebert &
Carpenter, 1992). “The depth of knowledge required by an expectation or in an assessment is related to the number of connections of concepts and ideas a student needs to make in order to produce a response, the level of reasoning, and the use of other self-monitoring processes. In addition, other factors influence the cognitive demands of performance, including the social or contextual requirements, the variety of representations students are expected to use (written, verbal, pictorial, and variations within each), and the requirements for transfer and generalization to new situations. (Webb, 2007).

**Content Validity and Alignment**

Content validity generally refers to the degree to which a test approximately represents the content domain it is intended to measure (Martone & Sireci, 2009). With respect to a content validity study, there are at least four potential aspects: domain definition, domain representation, domain relevance, and appropriateness of the test construction procedures (Sireci, 1998a, 1998b). Domain definition is the process used to define operationally the content domain tested. This is derived from the state-established curriculum frameworks, which for this study is the LOTE Checkpoint B syllabus. The domain representation refers to the degree to which the test represents and adequately measures all facets of the intended content domain. The four domains of content validity-- domain definition, domain representation, domain relevance, and appropriateness of the test construction procedures--in large measure mirror the four alignment criteria, which are categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation. The relevance of the domain to the content universe and the test construction procedures are
implemented on the policy level. Content validity can be supported if there were strong quality control measures in place that correspond to the domains.

Lane (1999) outlined procedures for evaluating the validity of assessments designed to measure students’ mastery of state academic standards. According to Lane, two forms of evidence are pertinent to determine the validity of these assessments: (a) the extent to which the state assessment reflects the state’s academic standards and (b) the extent to which the curriculum offered to students reflects the academic standards.

**Scale of Agreement**

The three-point scale used to describe cognitive alignment (depth-of-knowledge consistency), as defined by Webb, ranges from full to insufficient and is defined as follows:

- **Full** - For each major topic, the most cognitively demanding expected performance for all students is comparable to the most cognitively demanding assessment activity taken by all students.

- **Acceptable** - For nearly all major topics, the most cognitively demanding expected performance for all students is comparable to or can be inferred from the most cognitively demanding assessment activity taken by all students.

- **Insufficient** - Students can be judged as performing at an acceptable level on the assessments without having to demonstrate for any topic the attainment of the most cognitively demanding expected performance for all students.
Theoretical Framework: Constructive Alignment and Assessments

Ference Marton and Roger Saljö (1976) demonstrated that students learn not what teachers think they should learn but what students perceive the task demands of them (Biggs, 1996). In fact, learning takes place internally, as each student interacts with the material and constructs personal meaning, identifies rules and principles, and reviews the experiences. Some have called this the experienced curriculum (Boschee et al., 2008). The experienced curriculum is based on the concept of constructive learning, and it suggests that each student learns something different from the same lesson as he/she applies filters that are based on prior knowledge and social experiences (sometimes referred to as social constructivism). The term also implies that learning is created less by the teacher’s transmission and more by construction through the learner’s activity. It is not what the teacher does with the content but what the learner does with the content that is of more importance when it comes to achieving the learning objectives (Hattie, Biggs, & Purdie, 1996). This leads to the question “How does the teacher ensure that each student learns the same concept?” The current education policy view put forth by bureaucrats is that standardized learning requires constructive alignment, standardized curricula, and standardized testing or assessment (see Figure 1).
<table>
<thead>
<tr>
<th><strong>Teaching/Learning Activities</strong></th>
<th><strong>Curriculum Objectives</strong></th>
<th><strong>Assessment Tasks</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Designed to elicit desired verbs. May be:</td>
<td>Expressed as verbs students have to enact</td>
<td>Evaluate how well the target verbs are elicited and deployed in context.</td>
</tr>
<tr>
<td>Teacher Controlled</td>
<td>A Reflect Hypothesize, generate</td>
<td>The highest level verb To be clearly manifested becomes the final grade A, B, C, etc.</td>
</tr>
<tr>
<td>Peer Controlled</td>
<td>B Apply to ‘far’ domains Relate to principles</td>
<td></td>
</tr>
<tr>
<td>Student Controlled</td>
<td>C Apply to ‘near’ domains Analyze, compare Explain, solve Understand main ideas</td>
<td></td>
</tr>
<tr>
<td>As best suits the context</td>
<td>D Elaborate Classify Cover topics Describe</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 1.* Biggs Structure of Observed Learning Outcomes (SOLO Taxonomy).

Curriculum alignment encompasses the alignment of content, learning objectives, teaching method, learning activities, assessment, and evaluation with the expectations for
learning outcomes (Biggs, 1996). Biggs calls this system constructive alignment and elaborates that constructive alignment “represent a marriage between a constructivist understanding of the nature of learning, and an aligned design for teaching.”

**Constructivism**

In the education literature, constructivism is represented as a descriptive method of learning, teaching, education, cognition, personal knowledge, scientific knowledge, and a world view (Matthews, 2002). It is grounded by the twin principles of constructivism in learning and alignment in teaching. Therefore, the challenge is to leave room for each student to engage in constructivist learning while providing students with feedback and verifying through testing, that the expectations for learning outcomes are achieved.

This process of verifying that individual students learn the content leads to the debate of whether assessment should be norm referenced; i.e., based on a normal distribution and graded on a curve, or criterion referenced; i.e., passing grades are fixed and performance is measured against predetermined criteria. (Biggs, 1996). Biggs posited that norm-referenced assessments are judgments about people; criterion-referenced assessments are judgments about performance. Popham (1978) wrote that “For the purposes of instruction or evaluation, norm-referenced achievement tests are essentially worthless” (p. 6). Baker, Freeman, and Clayton (1991) questioned the viability of using standardized norm-referenced tests to assess either the improvement of an individual's education or the impact of systemic education reform. “Criterion-referenced tests [on the other hand], if properly fashioned, can be of enormous utility to instructors and evaluators” (Popham, 1978, p. 6). Criterion-referenced tests are used to ascertain an
individual’s status with respect to well-defined behavioral domain (Glasser, 1993).
Criterion-referenced tests provide explicit information as to what the individual can or cannot do. Criterion-referenced tests usually have cut scores. However, when levels define performance and terms such as proficient, advanced, and such are used, criterion-referenced tests offer little improvement over norm-referenced tests. “Proficient is merely an arbitrary point on a continuum of performance; it does not indicate mastery of a discrete set of skills” (Koretz, 2008, p. 29). To obtain reliable information about which students are at proficient levels as defined by state statutes, one needs test items that discriminate among students whose mastery is near the level of proficient (Koretz, 2008).

Rotherham (2006) and Kane (2010) wrote that “most high-stakes testing programs employ decision rules in which cut scores play a major role” (Rotherham, 2006, p. 3; Kane, 2010, p. 12). “Traditionally, the cut scores were determined by analyzing the utilities (or consequences) associated with the adoption of different cut scores (Cronbach & Gleser, 1965), where the utilities were defined in terms of such outcomes as the productivity of workers hired or the performances of students placed into different courses” (Kane, 2010, p. 12)

More recently, judgmental standard-setting studies have been designed to identify reasonable cut scores for educational assessments. The participants in a judgmental standard setting study develop (or are given) a description of the level of competence needed for some purpose (e.g., minimal competence for licensure decisions, or basic, proficient, and advanced for some accountability programs), which I call the performance standard, and a corresponding point on the score scale, which I call the cut score (Kane, 1994). The performance standard, or achievement level, is a verbal specification of some
level of achievement. Candidates with scores above the cut score have presumably met
the performance standard and are assigned to the higher category. Candidates with scores
below the cut score have presumably not met the standard and are assigned to a different,
lower category. (Kane, 2010, p. 12).

The LOTE syllabus and standards appear to be grounded in the constructivist
theory. Both are based on the need for students to create meaning with the content and
the real world. For example, tasks in Section 1 are related to real world experiences a
person may have when visiting a Spanish-speaking country. The Spanish Comprehensive
Regents examination is a criterion-referenced test. The criterion is presented in the New
York State syllabus. The passing grade or cut score is 65. With cut scores, even the
smallest amount of the standard error of measurement has consequences on students and
the pass/fail rate (Tienken, 2011).

The Communicative Model

Modern Language Examinations in NYS stress the communicative model versus
the traditional grammar-driven model. Communication is about making personal meaning
out of content. If there is no meaning, there is no communication. One of the learning
goals of the Modern Language program in New York State is that students will be able to
read print in the target language and make meaning of it. When listening to a speaker of
the target language, students will construct and understand the intended meaning and
respond appropriately. The meaning that they make will be understood by others and be
consistent with the original. Students also need to be able to respond to the subtleties that
native speakers communicate in terms of tone of voice and physical gestures within the
cultural complexities that impact communicating a coherent message. One example is
that in the Spanish culture, people greet each other before beginning a discussion. “Good morning, where is the bathroom?” “Buenas días, señor, ¿Dónde está el baño?” Although this might seem very logical oftentimes, second-language learners forget to include expressions of courtesy and greetings as well as the use of usted or vosotros in verb conjugations as a mark of respect.

Generally, communication is necessary and purposeful since most often there is an information gap where one person has the information and the other does not. Why would I pay attention to your answer if I already know the answer? This concept can be skewed in a second language classroom where the prompts are artificial. For example, the question “Are you wearing a shirt?” is nonsensical when clearly the person is standing in front of you and is fully dressed. However, in a different context (e.g., for someone getting ready to go to a party) the question “Are you wearing a shirt or a blouse?” would be relevant. Communication is unrehearsed and spontaneous. It always involves meaning, and it is always purposeful.

Of import to educators is how the seemingly competing objectives of constructivist learning and large-scale standardized assessment might be brought together. On one hand there is the necessity for students to construct meaning, which is aided by coherent vertical content alignment while, on the other hand, there are “standardized tests, which are uniform, and all learners face the same tasks administered in the same manner and scored the same way” (Koretz, 2008).
Curriculum Alignment

Curriculum alignment (Gamoran et al. & Webb, 1997; Cohen, 1997), has noticeable impact on student achievement as learning goals are explicit and made clear, and students perform better when they have had the opportunity to learn the material being assessed (Porter, 1997; Elliott & Roach, 2007; Boykin & Noguera, 2010).

There are three aspects that are characteristics of curriculum alignment: directionality, dimension, and level of analysis. Directionality refers to the direction in which alignment is examined, vertical and/or horizontal (Roach, Niebling, & Kurz, 2008). Dimension has two facets:

- Topical/Conceptual Knowledge
- Cognitive Complexity/Demand, Emphasis.

The level of analysis can either be coarse-grained or fine-grained.

![Alignment Diagram](image)

**Figure 2.** Characteristics of Alignment (Adapted from Niebling, Roach, Rahn-Blakeslee, 2008).

Curriculum maps draw explicit connections between learning benchmarks, learning experiences, and assessment tools. They help ensure that various aspects of
students’ learning experiences are aligned with one another; i.e., internal program alignment (Drake & Burns, 2004). They also help align learning benchmarks with mandated standards; i.e., external program alignment. An idea pioneered by Hausman in the 1970s, curriculum mapping has become a common practice in K-12 and post-secondary education (Willett, 2008).

Platform unity, alignment of classroom instruction, standards, and assessment are based on the Principles of Performance Instruction, and are a way to uphold curriculum alignment. The integration of planning and evaluating is often neglected in traditional approaches to instruction. Performance Instruction holds that course content should be planned, delivered, and evaluated consistently to assure unity. Test creation, for example, needs to be related to content. The domain level at which content is planned becomes the basis for creating test items, with content planning and testing at the same domain level to assure unity. (Leitzel & Vogler, 1994)

Content analyses are used together with graphic displays to make clear to policymakers, practitioners, and researchers the “content messages” of content standards and assessments. Information from content analyses of assessments, standards, and instruction has been used as the backbone of information for teacher professional development. An alignment index calculated on data from the content analyses is used as a teacher-level dependent variable in a number of program evaluations and research studies (Porter, 2002). Over 30 states and several big city school districts are using data from the content analysis procedures for professional development and to improve the quality of assessment instruments.
Alignment and Cognitive Complexity

In 1956, Benjamin Bloom headed a group of educational psychologists who developed a classification of levels of intellectual behavior important in learning. Bloom's Taxonomy is a multi-tiered model of classifying thinking according to six cognitive levels of complexity.

Bloom’s Taxonomy is one of the first systematic classifications of the processes of thinking and learning. The cumulative hierarchical framework consisting of six categories, each requiring achievement of the prior skill or ability before the next more complex one, remains easy to understand. Out of necessity, teachers must measure their students’ ability and plan how to achieve learning benchmarks. Accurately doing so requires a classification of levels of intellectual behavior important in learning. Bloom's Taxonomy provides the measurement tool for thinking. During the 1990s a new group of cognitive psychologists, led by Lorin Anderson (a former student of Bloom), updated the taxonomy to reflect relevance to 21st century work.
Webb (1997) defines curricular alignment as the extent to which and how well curricular categories and the elements (e.g., content standards, instructional content and assessment practices) work together to guide instruction and ultimately, facilitate and enhance student learning.

Depth of Knowledge (DOK) is the degree of understanding a student needs to respond to an assessment item. There are four DOK progressive levels: recall, skills, strategic thinking, and extended thinking.

- Level 1 - Recall: activities require the recall of information such as a fact, a definition, following steps in a simple procedure, performing simple
calculations, applying a formula, and making measurements. Students answering a Level 1 question either know the answer or they do not.

Action Verbs for use in creating Level 1 activities include the following: Arrange, Calculate, Define, Draw, Identify, Illustrate, Label, List, Match, Measure, Memorize, Name, Quote, Recognize, Repeat, Recall, Recite, Report, State, Tabulate, Tell. Also useful for creating Level 1 activities are the questions “Who, What, When, Where, Why.”

- Level 2 - Skill/Concept: activities require connecting recalled information and making decisions about problem solving or procedures. Level 2 activities require students to interpret and develop relationships between concepts.

Action verbs for use in creating Level 2 activities include the following: Cause/Effect, Classify, Categorize, Collect & Display, Compare, Construct, Distinguish, Graph, Infer, Identify Patterns, Interpret, Make Observations, Modify, Organize, Predict, Relate, Separate, Show, Summarize, Use Context Clues.

- Level 3 - Strategic Thinking: activities require reasoning, planning, using evidence, and in most cases involve the student’s explaining his or her thought processes. Typically, Level 3 activities have more than one correct response or approach to a problem.

Action Verbs for use in creating Level 3 activities include the following: Apprise, Assess, Construct, Compare, Critique, Cite Evidence, Compare, Draw Conclusions, Develop a Logical Argument, Differentiate, Explain Phenomena in Terms of Concepts, Formulate, Hypothesize, Investigate, Revise, Use Concepts to Solve Non-routine Problems

- Level 4 - Extended Thinking: questions require complex reasoning, planning, developing, most likely over an extended period of time (The extended time is
not a distinguishing factor if the work is repetitive and does not require applying significant conceptual understanding and higher-order thinking).

Action Verbs for use in creating Level 4 activities include the following: Apply Concepts, Analyze, Critique, Connect, Create, Design, Prove, Synthesize.


Figure 4. Webb's Depth of Knowledge.
Cognitive alignment can be determined using either Bloom’s Taxonomy or the Webb method of alignment. Webb’s work is used by educators in test item development as well as in alignment studies to determine the degree of alignment between the state standards (curriculum) and the tests used by states for accountability purposes.

**Webb Alignment Method**

The Webb alignment methodology is most appropriate for this study on content validity and alignment of the Spanish Regents to the NYS standards. In 1997, Webb conducted one of the first efforts of alignment of standards to assessment in the era of standards-based reform since the passage of the Goals 2000: Educate America Act (Webb, 1997). Webb (1999) developed the criteria for alignment in conjunction with the Council of Chief State School Officers to analyze alignment as it applied to mathematics and science in four states (Webb, 1999). Subsequently, he applied his alignment method to language arts standards and assessment in three states and to the Wisconsin standards in its alternative assessment for students with disabilities.

Webb found that the items on the state tests demanded a lower level of cognitive knowledge and skills than the standards expected. In four of the fourteen tests analyzed, the degree of consistency was quite high; in two cases, a fourth grade mathematics test and a sixth grade mathematics test, the match was 100 percent. But for the rest, the test items were rated at lower levels than the standards. Webb also found that the tests and standards tended to lack range-of-knowledge correspondence. To meet this criterion, tests must include items that measure at least half of the related objectives within a given standard. However, the study found that test items tended to cluster around a few objectives, leaving most of the
objectives un-assessed (Webb, 1999). The findings were similar in the studies of English language arts standards and tests. On the other hand, the Wisconsin Alternative Assessment for students with disabilities was relatively well aligned with state standards, (Rothman, 2003, p. 4).

In 2005, Webb used nine reviewers to evaluate the alignment between Michigan’s high school mathematics standards and six different assessments. Reviewers participated in a consensus process to determine the depth-of-knowledge levels of the Michigan high school objectives and then individually matched assessment items to objectives, goals, and standards and identified depth of knowledge of assessment items. The Michigan high school mathematics standards list six standards (e.g., patterns, relationships, and functions) with up to 18 objectives (e.g., analyze and generalize mathematical patterns including sequences, series and recursive patterns) for each one (Webb, N. M., Herman, & Webb, N. L., 2007). The depth-of-knowledge alignment findings are as follows:
Herman, Haertel, and the National Society for the Study of Education (2005) examined alignment between the Golden State Examination (GSE) in high school mathematics and the University of California Statement of Competencies in Mathematics (competencies expected for entering freshmen). Twenty reviewers individually rated the mathematics items of the GSE relative to the expectations identified in the University of California competency statement, identifying item features related to content and depth of knowledge (as well as an additional item feature, centrality). The University of California Statement of Competencies in Mathematics lists six content categories (e.g., variables, equations, and algebraic expressions) with up to 10 specific topics considered essential for entering college freshmen (e.g., solutions of linear equations and inequalities) in each
content category.

![Figure 6. Range-of-Knowledge Correspondence: Golden State Exam in Mathematics.](image)

Webb (2006) applied his model to evaluate alignment of math standards and assessments for Wisconsin for Grades 3-8 and Grade 10. Eight reviewers (six from Wisconsin and two 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 the training of reviewers. The reviewers were trained in the use of the four levels of depth of knowledge criteria by focusing on their definitions and examples. The
entire group of reviewers was involved in determining the depth of knowledge of the objectives. Individual rating of the items followed. 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 an 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 (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. An 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. One year later, Herman, Webb, & Zuniga (2007) conducted a case study on Measurement Issues in the Alignment of Standards and Assessments in which reliability data and inter-rater reliability is discussed. A read-behind procedure was advocated and the study found no correlation between the number of coders and accuracy of rating.

Webb (2009) directed an alignment review to evaluate the alignment of the Spring 2009 and Summer 2009 end-of-course test forms for English II, Algebra I, and Biology of the Missouri Course-Level Expectations. Using the Webb alignment method, reviewers rated individual test items on the cognitive complexity and the content assessed relative to the Missouri Course- Level Expectations. Webb (2009) used all four alignment
measures: (1) categorical concurrence - to determine the degree of overall content coverage by the assessment for each content strand, (2) range-of-knowledge representation - to indicate the specific content expectations assessed within each strand, (3) balance-of-knowledge representation - to provide a statistical index reflecting the distribution of assessed content within each strand (how evenly the content is assessed), and (4) depth-of-knowledge (DOK) consistency - to compare the cognitive complexity ratings of each content standard.

Webb (2009) found that for English II, the DOK alignment of the items to the standards was less than half. In fact, 52% were assessed at the lower levels. Similarly for Algebra, less than 50% of items were aligned to the content standards. In Biology as well, less than 50% of the DOK of items were aligned to the standards at DOK 1 and DOK 2 (Taylor, Koger, Koger, & Thacker, 2009).

The Common Core State Learning Standards were released in 2010. “In June 2010, the Council of Chief State School Officers (CCSSO) convened 35 specialists in Math and English Language Arts and Reading from 18 states to conduct a content analysis of the Common Core State Standards using the Content Frameworks of the SEC methodology” (Porter, McMaken, Hwang, & Yang, 2011, p. 105). The consortiums are in the process of developing the national standardized assessments; therefore, the Common Core State Standards were matched to the current Survey of Enacted Curriculum to identify the teaching learning gaps between the two standards. This is important as school systems across the United States prepare to switch from state standards to common (national) standards. For mathematics, the Common Core State Standards represent a modest shift toward higher levels of cognitive demand than are
currently represented in state standards. Of course, state standards vary considerably so that these differences would vary across states. In the CCLS, English Language Arts and Reading section, there is a greater emphasis on analysis and a decrease in emphasis on literal comprehension (Porter et al., 2011). There is an increase in emphasis on language study at the fine-grain level to “perform procedures/explain” than on higher levels of cognitive demand. Neither the Common Core nor the aggregate state standards are focused in their call for work on comprehension. Both cover most topics and most levels of cognitive demand (Porter et al., 2011).

Alignment Methodology

There are three widely used methods for undertaking alignment studies. Although they are different in their approach, they do contain similarities, which build on earlier notions of content validity. The alignment method to be used is determined by the goals of the study, with each method containing its own strengths and limitations (Porter, 1993). The three most common alignment methods are the Webb, Achieve, and Surveys of Enacted Curriculum.

Webb’s methodology of alignment focuses on assessments and standards. Each of the four dimensions and six subcategories provide a thorough understanding of alignment between the standard and the assessment. Webb (2005) wrote the following in Issues Relating to Alignment:

In 1998, the newly developed alignment process was used for the first time to analyze, with the cooperation of CCSSO, the alignment of curriculum standards and assessment of four states. Four to five reviewers coded the depth-of-knowledge (DOK) levels (cognitive complexity or demand and emphasis) of
standards and the assessment items using paper-and-pencil forms. These data were hand-entered into an Excel file and then analyzed using procedures developed with the help of John Smithson. Over the next two years, the alignment process was refined and used to conduct alignment analyses in additional states. The definitions for the depth-of-knowledge (DOK) levels for four content areas (reading and writing, mathematics, science, and social studies) were written and refined after each analysis (p. 2).

The Achieve Methodology, developed by Rothman and colleagues (2002) is an alignment protocol specific to English Language Arts, mathematics, and science. Unlike Webb’s methodology, Achieve does not have clear cut-off criteria to determine acceptable level of alignment. An expert panel independently conducts the alignment and the analysis, and the study is presented as a comprehensive report to the state.

Porter and Smithson (2001) developed the Surveys of Enacted Curriculum (SEC) methodology, to aid classroom teachers to see the connection between what is taught and what is assessed. The SEC methodology provides a snapshot of practice during a period of time, which is useful in determining the extent to which teaching reflects standards and assessments (Blank, Porter, & Smithson, 2001). The SEC model has two basic domains; namely, content match and cognitive demand. 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 includes five categories of cognitive demand: memorize, perform procedures,
communicate understanding, solve non-routine problems, and conjecture/generalize/prove. 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 four-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).

The SEC (Survey of Enacted Curriculum) method includes a measure of instructional content. Porter and Smithson (2002) emphasized the importance of including an instructional content component because it serves as an intervening variable when looking at student achievement gains because of standards-based reform. Through surveys, teachers code the instructional content as they think about a pre-selected target class during a specified period of time. Then the teachers estimate the emphasis allotted to that topic for each of the cognitive areas. This is then summed up to determine the proportion of each topic relative to the total instructional time (Martone, & Sireci, 2009, p. 1345).

Alignment is complex and comprised of two dimensions, topical conceptual knowledge and cognitive complexity/demand emphasis. Researchers use these two
dimensional frameworks to approach alignment studies regardless of the methodology used. Topical or conceptual knowledge deals with the skills and knowledge students are supposed to learn. These topics are matched to the standards, and cognitive complexity describes how much and what kinds of “thinking” are called for in the syllabus documentation and assessments. The panel coded and matched the items of the Languages Other Than English Comprehensive Spanish Regents Examination to the Webb’s Depth of Knowledge cognitive framework. I calculated the percentage of alignment to determine the level agreement (Webb, 2002; Herman & Webb, 2007).

Cognitive complexity involves content match and depth match. Depth alignment refers to the match between the cognitive complexity of the knowledge/skill prescribed by the standards and the cognitive complexity required by the assessment item/task (Webb, 1997, 1999).

Validity

Valid educational assessments require significant overlap between the content of the assessment questions and the underlying curriculum standards measured to ensure that all inferences made on the basis of test scores are defensible. Validity is the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of the tests (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 9) The authors of the NCLB Act proposed uses of tests to include the evaluation of students’ current academic proficiency in language arts, mathematics, and science and their progress with regard to the state-defined standards or curricular frameworks. Hence, evaluation tests for such purposes involve the consideration of both the content of
the test and the state-defined frameworks. Because test alignment studies cover the aspect of content and curricular frameworks, alignment studies provide validity evidence of the test and the curriculum. (Li & Sirecei, 2005). Evaluating how well the test items represent the domains specified in the test blueprint is one way to make inferences on curriculum and address concerns of curriculum narrowing (Lynn, 2000).

Test validity is separated into different categories: content validity, predictive validity, concurrent criterion-related validity, and construct validity. The key issues of test validity are the meaning, relevance, and utility of scores; the import or value implications of scores as a basis for action; and the functional worth of scores in terms of the social consequences of their use (Messick, 1990). Cronbach (1971) posited that what is to be validated is not the test or the observation device . . . but the inferences derived from the test scores. In fact, test validation is an empirical evaluation of the meaning and consequence of measurement. Test validity determines whether the curricular inferences made on the basis of test scores are accurate. Test validity also encompasses the implications of the test scores and indicates that the skills are facilitative of college skills (relevance), that the scores are useful to predict success in freshman year (utility), that there is no adverse impact against females or minority groups, and that there is no male-oriented item content or other sources of construct-irrelevant test variance, but rather that the test reflects authentic group differences in construct-relevant quantitative performance (Messick, 1990). Thus, validity carries with it implications of scores as the basis for action and social consequences. At its center is the concept that the test items are samples of the behavioral domain or item universe about which inferences are to be drawn and predications made.
Content Validity

Content validity refers to the degree to which a test approximately represents the content domain it is intended to measure. When a test is judged to have high content validity, its content is considered to be congruent with the testing purpose and with prevailing notions of the subject matter tested. (Martone & Sireci, 2009). Therefore, alignment is one facet of validity.

Content validity has four categories or domains: (1) domain sampling, (2) domain representation, (3) domain relevance and appropriateness of test construction procedures, and (4) domain clarity and the technical quality of test items (Sireci, 1998a, 1998b). Cronbach (1980) advised that the items sample rigorously from the domain of tasks deemed relevant to the test construct.

Ebel (1955), obtaining and reporting evidence of content validity--contributions to a symposium on content validity of non-factual tests in San Francisco--stated, “Only when the content of education is conceived as a set of goals to be attained, rather than a set of lessons to be studied or as a set a class activities to be carried out, is it educationally useful to seek content validity in a test.” He also posited that educational achievement tests are designed to measure the degree of attainment of the ultimate goals of instruction in a particular area. The more directly and completely and reliably a test measures the attainment of these goals, the greater is its content validity (Ebel, 1997).

Factors That Affect Content Validity

The Spanish Comprehensive Regents Examination has several of the tasks written in English. Table 1 illustrates the test section and the language in which it written, the
language of the questions, and the language of the performance tasks/answers. All of the
instructions in the test are written in English.

In Table 1, I provide an outline of the format of the New York State Spanish
Comprehensive Regents Examination. There are four sections. With the exception of
Part 1, each section contains subdivisions with items and student responses either in
English or the target language, Spanish.

Table 1

<table>
<thead>
<tr>
<th>Test section</th>
<th>Items</th>
<th>Student Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>Spanish</td>
</tr>
<tr>
<td>2A</td>
<td>Spanish</td>
<td>English</td>
</tr>
<tr>
<td>2B</td>
<td>Spanish</td>
<td>Spanish</td>
</tr>
<tr>
<td>3A</td>
<td>Spanish</td>
<td>Spanish</td>
</tr>
<tr>
<td>3B</td>
<td>Spanish</td>
<td>English</td>
</tr>
<tr>
<td>3C</td>
<td>Spanish</td>
<td>English</td>
</tr>
<tr>
<td>4A</td>
<td>English</td>
<td>Spanish</td>
</tr>
<tr>
<td>4B</td>
<td>English</td>
<td>Spanish</td>
</tr>
</tbody>
</table>

As shown in Table 1, students who are English dominant have an advantage when
taking the Spanish Comprehensive Regents Examination, as all of the instructions are
written in English. This lends itself to teachers teaching students how to recognize the
format of the test.

The second factor that can affect test score validity is that the test writer needs to
determine who the test takers are, what their language levels are, the purpose of the test in
relation to the institution, the uses of the results of the test, and the time constraints of the
test (Alderson, Clapham, & Wall, 1995). The test takers in NYS are students in Grades
10, 11, and 12 and the stated goal of the test developers is to measure how much of the
content presented on the test the students know. The results of the test provide a mechanism for state education bureaucrats to hold school personnel accountable for the passing rate of students on the examination. Bureaucrats also use the results to determine whether students receive an Advanced Regents diploma. Parts 2, 3, and 4 are three hours in duration. As a result of the high stakes attached to the assessment, it is therefore essential that test developers understand the institutional objectives and the intended purpose of the specific test being created (Genesee & Upshur, 1996, De Benedetti, 1987). The Modern Languages Regents Examination measures the domains of speaking, listening, reading, writing, the issue of import that there is alignment of the instrument to the Checkpoint B syllabus.

Other factors that can affect test validity include but are not limited to the following:

- Inappropriate selection of constructs or measures
- Insufficient data collected to make valid conclusions
- Measurement done in too few contexts
- Measurement done with too few measurement variables
- Too great a variation in data
- Inadequate selection of sample population for pilot exam

Summary

Policymakers embarking on an effort to create a more effective system less prone to the drawbacks of simple test-based accountability cannot rely solely on alignment and should consider several additional steps: (1) redesigning external tests in other ways to minimize inflation, (2) setting attainable performance targets, (3) relying on multiple
measures, and (4) reestablishing a role for professional judgment. Developing more effective alternatives will take us beyond what is well established and will require innovation, experimentation, and rigorous evaluation (Koretz, 2005).

A “positive washback” effect occurs when the assessment procedures correspond to the course goals and objectives (Brown & Hudson, 1998, p. 668; Hamidi, 2010). Testing, when designed in a manner sensitive to the content area, can influence instruction. Tests can inform the teacher as to the next steps in instruction and inform the students of their achievement and deficits within a specified framework of skills.

“Positive washback benefits teachers, students, and administrators because it assumes that testing and curriculum designs are both based on clear course outcomes that are known to both students and teachers/testers” (Coombe, Folse, & Hubley, 2007, p. xxv). "To achieve the positive washback effect, Hasselgren (2000) maintained, that the tasks should promote good learning as well as assessment activities [in which] both pupils and teachers should develop their ability to assess, based on explicit criteria" (p. 262). Well-designed performance assessments can actually provide strong positive washback effects, especially if they are directly linked to a particular curriculum (Hamidi, 2010).
CHAPTER 3

METHODOLOGY

Design

I used a qualitative case study design for this study. The purpose of this research design was to outline the entire logical structure and process of inquiry of the research project, including framing the research questions and data analysis. The intent of the study, which provides the context for the research design, was an analysis of the content validity and the cognitive level of the Comprehensive Spanish Regents Examination and the determination of the alignment of the Spanish Regents assessment with the New York State Checkpoint B syllabus as required by No Child Left Behind (2001). This intent, therefore, helped the researcher devise a plan to investigate the subject matter to ensure that the data collected provide evidence in support of the research questions under consideration. The case study is the research design type utilized in this qualitative study.

A qualitative case study is most appropriate for a work such as this. The subject matter for the study lends itself to a qualitative analysis, which is justified on three levels. First, the investigator seeks to explore human behavior in the context of the Comprehensive Regents Examination in New York State by collecting appropriate data and analyzing and interpreting them through coding and matching. This is different from a quantitative study that collects, analyzes, and interprets data “by observing figures and numerical patterns” (Boodhoo & Purmessur, 2009). Second, qualitative analysis is appropriate for case studies, (Merriam, 2009), and this work is a case study. Third, the qualitative method generates rich data, has a flexible structure of inquiry, and supports the ideals of inductive reasoning, creating meaning, and the importance of rendering
solutions to complex situations (Merriam, 2009). It is worthy of note nonetheless that the equating of “a particular research design with either quantitative or qualitative methods” is not a position that is universally embraced in the literature (De Vaus, 2003). Some scholars argue that it is erroneous to do so because a “quantitative/qualitative distinction for case studies” is irrelevant. Yin (1993) noted that “People have thought that the case study method required them to embrace these data collection methods [qualitative method, ethnography, participant observation] . . . On the contrary, the method does not imply any particular form of data collection--which can be qualitative or quantitative” (Yin, 1993, p. 32 as cited in De Vaus, 2003, p. 11).

There are several methodologies that can be used to ascertain the degree of alignment and arrive at findings; however, it is important to consider what it takes to reliably capture the curriculum alignment data. At the heart of measuring curriculum alignment is the comparison of two curricular elements, the assessment and the standards, and to determine how similar they are to each other.

**Methods**

I used the framework of Webb alignment protocol and read-behind method of coding to conduct the alignment measuring process. The Webb Alignment Model has four dimensions (Webb, 2005): categorical concurrence, depth-of-knowledge consistency (DOK), range-of-knowledge correspondence, and balance representation. Webb’s Depth of Knowledge framework has four levels: recall, basic application, strategic thinking, extended thinking (Webb, 2005). The Webb protocol also has specific set criteria for what constitutes a good enough match. When the criteria are not met, the coding can be
traced to specific standards and objectives, and the researcher can determine which areas need to be further addressed in order to meet the criteria.

I created an Excel spreadsheet to record the findings of each coder. The headings were documented in an excel spreadsheet under the headings, Year/month/day, Question number, Depth of Knowledge level, Depth of Knowledge description, Depth of Knowledge notes, Category number, Category/Theme, Topic, Notes, Checkpoint, Standard, Notes.

I trained coders on the DOK framework and the coding process, and each coder independently assigned single DOK levels to objectives and to individual assessment items. If there was rater disagreement, it was generally acceptable to choose the average of the two levels under consideration (Webb, 2005). In the final analysis, DOK is considered matched if at least 65% of the items are sufficiently aligned. The same procedure was used for the alignment of the items to the state syllabus. Through consistent coding of items, I arrived at the findings for the research questions.

**Research Questions**

I used the Webb alignment protocol and a read-behind method of coding to arrive at findings. The study was guided by the following questions:

1. What is the categorical concurrence and proportion of the overlap of Part 2 of the New York State Comprehensive Spanish Regents Examination 2006-2011? What is the depth-of-knowledge consistency as defined by Webb’s Depth of Knowledge of the New York State Comprehensive Spanish Regents Examination items for the years 2006-2011?
2. What is the depth-of-knowledge consistency as defined by Webb’s Depth of Knowledge of the New York State Comprehensive Spanish Regents Examination items for the years 2006-2011?

3. What is the range-of-knowledge correspondence, defined as the number of content objectives of the New York State Checkpoint B syllabus assessed by items Part 2 of the New York State Comprehensive Spanish Regents Examination 2006-2011? The range of knowledge is defined as the number of content objectives of the New York State Comprehensive Spanish Regents Examination 2006-2011.

**Researcher’s Role**

The researcher is the supervisor of the writing of the New York City Proficiency Examinations in French and Latin and the supervisor of the writing of the New York City Comprehensive Regents Examinations in Latin, French, and Spanish.

Before 2010, the New York State Department of Education invited educators to submit questions for the LOTE examinations. NYSED curriculum specialists selected schools across the state to pilot the test questions and encouraged feedback from the field after the administration of each test on the assessment items. None of the psychometric data from these examinations were released to the public. When NYSED personnel discontinued the writing and distribution of LOTE examinations, citing lack of budgetary resources, school districts were forced to assume the responsibility for creating these tests, which the law required for students to graduate with an Advanced Regents diploma. I was selected to supervise the writing of these LOTE examinations for the district of New York City by the NYC Office of Assessment and Accountability. This study came
from the need to determine the alignment levels of the State testing instruments so that new district tests can be created in accordance to the same alignment guidelines.

To avoid issues of bias, the years during which I was the supervising writer (assessments for year 2011 onwards) were not selected as part of this study. This study seeks to understand the content validity and alignment of the Spanish Regents Examination in order to create a blueprint with which to develop subsequent Modern Languages Examinations.

I am also on the panel that creates the New York State School Building Leader Assessment. I am part of a panel that creates the benchmark of the assessments, the items, and review psychometric data from initial piloting of the examination. I have received extensive training in standards alignment and alignment of items to the standards and syllabus.

I am also on the panel that is in the process of creating the National World Languages Examination.

**Data Collection**

I entered the NYSED website into my browser (http://www.nysedregents.org/Spanish/).

I downloaded and printed the January and June Spanish Comprehensive Regents Examinations for the years 2006-2011. I also printed the teacher’s dictation and answer key. (Miles & Huberman, 1994).

I created an Excel spreadsheet to log all findings. All raters recorded their findings digitally onto the Excel spreadsheet. The spreadsheet contained twelve horizontal cells, A-L. The headings from left to right began with A - Year/mth/day and
ended with notes on the right - L. Under the heading A - Year/mth/day, we wrote the date of the examination administration as indicated by the question booklet and dictation and answer key booklet. The next heading, B - Q#, question number, was used to document the number of the item being coded. The next two headings pertained to the depth of knowledge consistency and were C - DOK level, D - DOK description, and E - DOK notes. Under the heading DOK level, the coders recorded the DOK levels, 1-4 of the assessment item. Under the heading DOK notes, the coders documented how they arrived at the DOK rating. The next heading, F - category was in the domain of range of knowledge. The coders matched the items to the Checkpoint B syllabus of which there are fifteen categories; in the next column, G, the category and theme were documented. Next, the H topic, which is a subcategory of category and theme in the checkpoint syllabus, was documented. The next column, I, contained a brief synopsis of the items. Then the items were matched to the J-appropriate LOTE Checkpoint A, B, or C. The next headings fall under the domain of categorical concurrence. The items were matched to the K standards. The last column documented the L notes of the coders regarding the standard/item match.

The balance of knowledge is a delimitation and this does not dilute the study. Balance-of-knowledge representation provides a statistical index reflecting the distribution of assessed content within each strand; that is, how evenly the content is assessed and whether the assessment measures the content objectives equitably within each domain. The data are available to calculate the balance-of-knowledge representation in a future study. The raw data to calculate the balance-of-knowledge representation is attached in Appendix D.
The Testing Instrument

The LOTE syllabus was mandated to school districts since the mid 1980s. There are several factors that influenced the stability and caused changes to the examination. After the passage of No Child Left Behind in 2001, there was a need to measure yearly progress in all core subjects. Consequently, Regents Examinations were reviewed and some were revised. In New York State, Languages Other Than English is considered a core subject. All new Regents Examinations undergo a validation process, which lasts three years. NYSED redesigned the LOTE examination in 2003. The examination was divided into four sections. In Part 1, the speaking rubric was clarified and the writing rubric was introduced. The reason for the changes to the speaking and writing rubric is that implementation would standardize grading and reduce rater subjectivity. The Spanish Comprehensive Regents Examinations reviewed for this study have the revised format as is now mandated by NYSED.

The examinations from the years 2006-2010 were studied. There are two examinations in each given year, one administered in January and the other administered in June. The total number of examinations reviewed was 10. The test instruments can be located on the NYSED website (http://www.nysedregents.org/Spanish/).

The Spanish Comprehensive Regents Examination is comprised of four sections. The test is administered in two parts. Part 1, the speaking tasks, is administered as much as three months prior, and up to five days prior to Part 2, which is comprised of listening, reading, and writing. NYSED has provided 500 speaking tasks, created in 1999, of which the teacher or administrator chooses 60 tasks for Part 1 of the school’s
examination. The other tasks are kept secure for future tests administrations. Students select two tasks at random from the 60, and this portion of the examination is conducted in a one-to-one setting with the teacher. The tasks require that either the teacher or the student initiate the conversation. Ideally, a student would perform one of each. As explained in the delimitations, Part 1 of the Spanish Comprehensive Regents Examination is beyond the scope of this study.

Part 2 of the Comprehensive Spanish Regents Examination, has 33 questions. The instructions on the cover page are as follows:

This booklet contains Parts 2 through 4 (76 credits) of this examination. Your performance on Part 1, Speaking (24 credits), has been evaluated prior to the date of this written examination. The answers to the questions on this examination are to be written in the separate answer booklet. Be sure to fill in the heading on the front of your answer booklet. When you have completed the examination, you must sign the statement printed at the end of the answer booklet, indicating that you had no unlawful knowledge of the questions or answers prior to the examination and that you have neither given nor received assistance in answering any of the questions during the examination. Your answer booklet cannot be accepted if you fail to sign this declaration.

Part 2 begins with Section 2, Listening Comprehension. The subject teacher reads a brief background of the passage in English, then reads a short passage in Spanish. In the student’s question booklet are the question and four multiple-choice responses. In their answer booklet, students write A, B, C, or D—whichever they think is the correct
answer. There are nine questions in this section. Both the questions and the answers are in English. The passage is in the target language, Spanish.

The instructions to the student are as follows:

Directions (1–9): For each question, you will hear some background information in English once. Then you will hear a passage in Spanish twice and a question in English once. After you have heard the question, the teacher will pause while you read the question and the four suggested answers in your test booklet. Choose the best-suggested answer and write its number in the space provided in your answer booklet. Base your answer on the content of the passage, only. [18]

Questions 10-15 of Part 2, Listening Comprehension, begin with a brief background in English read by the teacher. The passage is in Spanish as in the previous section and is read twice. A question is then asked. This question is in the target language, Spanish. The directions are as follows:

Directions (10–15): For each question, you will hear some background information in English once. Then you will hear a passage in Spanish twice and a question in Spanish once. After you have heard the question, the teacher will pause while you read the question and the four suggested answers in your test booklet. Choose the best-suggested answer and write its number in the space provided in your answer booklet. Base your answer on the content of the passage, only. [12]

The third section is the Reading Comprehension section. Section 3 is divided into three parts A, B, and C. The questions are numbered 16-30.
In Section 3A, the questions are numbered 16-20. There is a reading passage in Spanish. The questions and the answers are in Spanish. The instructions for Section 3, part A, are as follows:

Directions (16–20): After the following passage, there are five questions or incomplete statements in Spanish. For each, choose the word or expression that best answers the question or completes the statement according to the meaning of the passage, and write its number in the space provided in your answer booklet.

[10]

Part 3B, is the realia section. This section contains advertisements, clippings from newspapers and magazines that students are likely to encounter in print and online media. There are 5 realia questions. The realia are in Spanish; however, the prompts and questions are in English. The directions are as follows:

Directions (21–25): Below each of the following selections, there is either a question or an incomplete statement in English. For each, choose the word or expression that best answers the question or completes the statement according to the meaning of the selection, and write its number in the space provided in your answer booklet. [10]

Part 3C continues the Reading Comprehension section. There is a passage in Spanish, and the questions and answers are in English. There are 5 questions. The directions are as follows:

Directions (26–30): After the following passage, there are five questions or incomplete statements in English. For each, choose the word or expression that best answers the question or completes the statement according to the meaning of
the passage, and write its number in the space provided in your answer booklet.

Section 4 is the Essay section. In Section 4, the final section, there are three essay-writing prompts. Two are based on writing prompts, and the final questions have a picture as a prompt. The instructions are as follows:

Write your answers to Part 4 according to the directions below. Your answers must be written in your own words; no credit will be given for a response that is copied or substantially the same as material from other parts of this examination.

Directions (31–33): Choose two of the three writing tasks provided below. In your answer booklet, write your response to the two writing tasks you have chosen. For each question you have chosen, your answer should be written entirely in Spanish and should contain a minimum of 100 words. Place names and brand names written in Spanish count as one word. Contractions are also counted as one word. Salutations and closings, as well as commonly used abbreviations in Spanish, are included in the word count. Numbers, unless written as words, and names of people are not counted as words. You must satisfy the purpose of the task. Be sure to organize your response and to include a beginning, middle, and ending. The sentence structure and/or expressions used should be connected logically and should demonstrate a wide range of vocabulary with minimal repetition.

For Question Number 33, there is picture prompt. The instructions are as follows:

In Spanish, write a story about the situation shown in the picture below. It must be
a story relating to the picture, not a description of the picture. Do not write a dialogue.

After Question 33, there is the essay-grading rubric.

Table 2 is the NYS Spanish Comprehensive Regents Essay Grading Rubric provided to students. Table 2 is listed after the final item in the question booklet.

Table 2

*NYS Spanish Comprehensive Regents Essay Grading Rubric for Student*

<table>
<thead>
<tr>
<th>Dimension</th>
<th>A response receiving maximum credit:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose/Task</td>
<td>Accomplishes the task; includes many details that are clearly connected to the development of the task, but there may be minor irrelevancies.</td>
</tr>
<tr>
<td>Organization</td>
<td>Exhibits a logical and coherent sequence throughout; provides a clear sense of a beginning, middle, and end. Makes smooth transitions between ideas.</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>Includes a wide variety of vocabulary that expands the topic, but there may be minor inaccuracies.</td>
</tr>
<tr>
<td>Structure/Conventions</td>
<td>Demonstrates a high degree of control of Checkpoint B (Regents level) structure/conventions:</td>
</tr>
<tr>
<td></td>
<td>subject-verb agreement</td>
</tr>
<tr>
<td></td>
<td>tense</td>
</tr>
<tr>
<td></td>
<td>noun-adjective agreement</td>
</tr>
<tr>
<td></td>
<td>correct word order</td>
</tr>
<tr>
<td></td>
<td>spelling/diacritical marks (e.g., accents)</td>
</tr>
<tr>
<td>Word Count</td>
<td>Contains 100 words or more.</td>
</tr>
</tbody>
</table>

In Part 2, texts gradually increase in length and cognitive complexity as defined by Webb’s depth-of-knowledge. However, as the level of cognitive complexity as defined by Webb’s Depth of knowledge goes up, the amount of points allocated goes down. According to the technical manual, questions are vertically scored with the exception of Part 4, which has a scaled score. In other words, each of the multiple-choice questions in Sections 2 and 3 are worth 2 points each. In Section 4, students write two
essays and demonstrate understanding of the language, its morphology, and syntax in the target language. This section is worth 8 points total, with each essay worth 4 points.

Table 3 displays the score conversion chart for the essay section, Part 3, of the NYS Spanish Comprehensive Regents Examination. NYSED provides this chart for the graders of the examination. Graders convert the raw score for each of the two essays the student is required to write, into a scale score as specified by NYSED.

Table 3

_Part 4 Score Conversion Chart. Listed on the Scoring Rubric for Graders of the Examination_

<table>
<thead>
<tr>
<th>Total Raw Score</th>
<th>17–18</th>
<th>15–16</th>
<th>13–14</th>
<th>11–12</th>
<th>8–10</th>
<th>6–7</th>
<th>4–5</th>
<th>2–3</th>
<th>0–1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Credits</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Testing instruments are released to the public on the NYSED website (http://www.nysedregents.org/Spanish/) up to the 2011 testing year. Subsequent examinations are created by individual districts and are kept secure by the specific district assessment bureau.

**LOTE Standards**

The LOTE standards were instituted in 1996; however, of the four standards only two were officially adopted by the New York State Department of Education. While individual districts used all four standards, the fact that NYSED implemented only two of the standards meant that the assessments were designed to align to those two NYSED standards. The standards were published as _Standards for Foreign Language Learning, Preparing for the 21st Century_. The Modern Language Standards began with background
information on the federal policy, Goals 2000, which was the driving force behind the standards movement, followed by a paragraph that states the following:

The standards do not describe the current status of foreign language education in this country. While they reflect the best instructional practice, they do not describe what is being attained by the majority of foreign language students. The Standards for Foreign Language Learning will not be achieved overnight; rather, they provide a gauge against which to measure improvement in the years to come. The standards are not a curriculum guide. While they suggest the types of curricular experiences needed to enable students to achieve the standards, and support the ideal of extended sequences of study that begin in the elementary grades and continue through high school and beyond, they do not describe specific course content, nor recommended sequence of study. They must be used in conjunction with state and local standards and curriculum frameworks to determine the best approaches and reasonable expectations for the students in individual districts and schools.

These standards became widely known as the five C’s, as there are five categories beginning with the letter “C,” under which the standards are written. The categories are Communication, Culture, Connections, Comparisons, and Communities. Under the heading Communication, the standards are written as follows:

Communicate in a Language Other Than English

Standard 1.1: Students engage in conversations, provide and obtain information, express feelings and emotions, and exchange opinions.

Standard 1.2: Students understand and interpret written and spoken language on a
variety of topics.

Standard 1.3: Students present information, concepts, and ideas to an audience of listeners or readers on a variety of topics.

Under the heading Cultures, the standards are written as follows:

Gain knowledge and understanding of other cultures

Standard 2.1: Students demonstrate an understanding of the relationship between the practices and perspectives of the culture studied.

Standard 2.2: Students demonstrate an understanding of the relationship between the products and perspectives of the culture studied.

Under the heading Connections, the standards are written as follows:

Connect with other disciplines and acquire information.

Standard 3.1: Students reinforce and further their knowledge of other disciplines through the foreign language.

Standard 3.2: Students acquire information and recognize the distinctive viewpoints that are only available through the foreign language and its cultures.

Under the heading Comparisons, the standards are written as follows:

Develop insight into the nature of language and culture

Standard 4.1: Students demonstrate understanding of the nature of language through comparisons of the language studied and their own.

Standard 4.2: Students demonstrate understanding of the concept of culture through comparisons of the cultures studied and their own.

Under the heading Communities, the standards are written as follows:

Participate in multilingual communities at home and around the world
Standard 5.1: Students use the language both within and beyond the school setting.

Standard 5.2: Students show evidence of becoming life-long learners by using the language for personal enjoyment and enrichment.

The standards that were adopted and implemented by NYSED and used in the creation of all Modern Languages Comprehensive Regents Examinations are as follows:

Standard 1: Students will be able to use a language other than English for communication.

Key Idea: LISTENING AND SPEAKING are primary communicative goals in modern language learning. These skills are used for the purposes of socializing, providing and acquiring information, expressing personal feelings and opinions, and getting others to adopt a course of action.

Performance Indicators--Students will:

Checkpoint B

Comprehend messages and short conversations when listening to peers, familiar adults, and providers of public services either in face-to-face interactions or on the telephone.

Understand the main idea and some discrete information in television, radio, or live presentations.

Initiate and sustain conversations, face to face or on the phone, with native-speaking or more fluent individuals.

Select vocabulary appropriate to a range of topics, employ simple and complex sentences in present, past, and future time frames, and express details and nuances.
by using appropriate modifiers.

Exhibit spontaneity in their interactions, particularly when the topic is familiar, but often rely on familiar utterances.

Use repetition and circumlocution as well as gestures and other nonverbal cues to sustain conversation.

Key Idea: READING AND WRITING are used in Languages Other Than English for the purpose of socializing, providing and acquiring information, expressing personal feelings and opinions, and getting others to adopt a course of actions

Performance Indicators--Students will:

Checkpoint B

Read and comprehend materials written for native speakers when the topic and language are familiar.

Use cognates and contextual and visual cues to derive meaning from texts that contain unfamiliar words, expressions, and structures.

Read simple materials independently, but may have to guess at meanings of longer or more complex material.

Write short notes, uncomplicated personal and business letters, brief journals, and short reports.

Write brief analyses of more complex content when given the opportunity for organization and advance preparation, though errors may occur more frequently.

Produce written narratives and expressions of opinion about radio and television programs, newspaper and magazine articles, and selected stories, songs, and literature of the target language.
Standard 2: Effective communication involves meanings that go beyond words and require understanding of perceptions, gestures, folklore, and family and community dynamics. All of these elements can affect whether and how well a message is received.

Performance Indicators--Students will:

Checkpoint B

Exhibit more comprehensive knowledge of cultural traits and patterns.

Draw comparisons between societies.

Recognize that there are important linguistic and cultural variations among groups that speak the same target language.

Understand how words, body language, rituals and social interactions influence communication.

Data Analysis

Qualitative research often categorizes data into patterns as the primary basis for organizing and reporting results. Coding is an interpretive technique that both organizes the data and provides a means to introduce the interpretations. Each segment is labeled with a "code," usually a word or short phrase that suggests how the associated data segments inform the research objectives. There are many different ways of establishing validity of coding, including prolonged engagement, member check, interviewer corroboration, AND peer debriefing. Most of these methods were coined, or at least extensively described, by Lincoln and Guba (1985). For this qualitative study, the coders used a read-behind method, which provided an additional layer of quality control and external validity of the coding process. The coders completed one test per day in a
duration of six hours. Members discussed their findings after each item was completed. There was a debriefing session at the end of each day of coding regarding any outstanding issues. These issues were resolved before the session ended.

**Rating Procedure and Protocols**

The alignment coding process entailed three panel members rating alignment between the test items and Webb’s Categorical framework of depth of knowledge. According to Webb (2002), the alignment coding process is not designed to produce exact agreement between members of the expert panel. In fact, variance in ratings “are considered valid differences in opinion that are a result of a lack of clarity in how the objectives were written and/or the robustness of an item that may legitimately correspond to more than one objective” (p. 3).

**Rater Panel**

Content analysis by expert panels remains the primary technique for judging alignment between standards and assessments. A panel of two persons served as raters over a twelve-week period. Candidates were recommended by a curriculum specialist, with selections based on expertise and experience in teaching Spanish, familiarity with the New York State Spanish Comprehensive Regents Examination and NYS Syllabus and Standards, and instructional expertise. Raters were screened and provided with training. All raters are high school teachers of Spanish general education classes or a related university training program.

**Rater Protocols and Inter-rater Agreement**

The panel met for a two-day session, starting with an orientation, covering definitions and an overview of the alignment process and the objectives of the study.
Raters were given the rating scales and protocols to review several days prior to the rating session, and their first task was to reach consensus on the depth-of-knowledge rating, then match test items to one of the four DOK levels. The second task was to reach consensus on the standards rating, and the third and final task was to practice and discuss sample matching of the item on the assessment to the checkpoint syllabus.

Panel members used two standardized protocols to independently apply the alignment criteria. The protocols were derived from alignment protocols by Webb (2005) and Porter (2007).

The protocol applied Webb's Depth-of-Knowledge scale, representing four knowledge levels to each benchmark, with the ratings summed up by group discussion and consensus. Webb's scale is briefly described below, with raters receiving a more extensive version with examples:

- **Level I** requires students to receive or recite facts or to use simple skills or abilities, such as oral reading.
- **Level II** calls for some mental processing beyond recalling or reproducing a response, shows comprehension and processing of text.
- **Level III** represents deep knowledge beyond text, which may include explaining, generalizing, or connecting ideas.
- **Level IV** shows higher-order thinking plus deep knowledge, typically requiring extended time to assess.

To ensure inter-rater reliability, each rater recorded his or her findings. For each action by an individual rater, the team discussed the findings. When individual raters gave a different rating for a single item, the team discussed the findings. Observations...
were recorded under the heading notes on the Excel sheet. There was no opportunity, nor was it required, for reviewers to offer their opinions on either the quality of the standards or of the assessment activities/items, as the verification of the general quality of the standards and the assessments are beyond the scope of this study. All raters were knowledgeable about the testing instrument, Webb’s Depth of Knowledge theory. The multiple codings served as an added check for accuracy. (Armstrong, 1997).

**Data Analysis and Coding**

The alignment coding process entailed two panel members rating alignment between the test items and Webb’s Categorical Framework of Categorical Concurrence, Depth of Knowledge, and Range of Knowledge. The alignment analysis was also completed by the panel members who provided descriptive statistics for three of the four criteria underlying Webb’s Alignment Model: (a) categorical concurrence, (b) depth-of-knowledge consistency, and (c) range-of-knowledge correspondence. The fourth category, balance of representation, is beyond the scope of this study.

The two panel members recorded their findings on a digital Excel spreadsheet with headings as shown below:

The spreadsheet contained twelve horizontal cells, A-L. The headings from left to right began with A - Year/mth/day and ended with notes on the right, L. Under the heading A - Year/mth/day, we wrote the date of the examination administration as indicated by the question booklet and the dictation and answer key booklet. The next heading, B - Q#, question number, was used to document the number of the item being coded. The next two headings pertained to the depth-of-knowledge consistency and were C - DOK level, D - DOK description, and E - DOK notes. Under the heading DOK
level, the coders recorded the DOK levels, 1-4 of the assessment item. Under the heading DOK level, the panelist recorded which of the four categories of DOK matched the cognitive level of the item under review. The four categories of DOK are recall, basic application, strategic thinking, and extended thinking.

- **Level I** requires students to receive or recite facts or to use simple skills or abilities, such as oral reading.
- **Level II** calls for some mental processing beyond recalling or reproducing a response; shows comprehension and processing of text.
- **Level III** represents deep knowledge beyond text, which may include explaining, generalizing or connecting ideas.
- **Level IV** shows higher-order thinking plus deep knowledge, typically requiring extended time to assess.

Panelists recorded the precise DOK labels recall, basic application rather than levels to ensure more precise recordings than the roman numerals.

Under the heading DOK notes, the coders documented how they arrived at the DOK rating. The next heading, F - category, was in the domain of range of knowledge. The coders matched the items to the Checkpoint B syllabus, of which there are fifteen categories; in the next column, G, the category and theme were documented. Next the H topic, which is a subcategory of category and theme in the checkpoint syllabus, was documented.

Under the heading “category and theme,” the panelists coded the match of the item to the theme of the checkpoint syllabus. For example, all panelists coded the Item
Number 22 (January 25, 2006) to the category/theme of “Lodging.” Under the heading “Topic,” Item Number 22, was coded “Hotel.”

![Image](22.png)

**Figure 7.** *NYS Spanish Regents Examination, January 25, 2006, Question 22.*

The next column, I, contained a brief synopsis of the items. Then the items were matched to the J-appropriate LOTE Checkpoint A, B, or C. The next headings fall under the domain of categorical concurrence. The items were matched to the K standards. The last column documented the L notes of the coders regarding the standard/item match.

Under the heading Standards, the panelist coded the item to the NYS Learning Standards, of which there are two: communication and cultural understanding. For the assessment January 25, 2006, Item Number 22, all panelists coded the item to Standard Number 1, Checkpoint B. Standard Number 1, Checkpoint B, reads as follows:

Read and comprehend materials written for native speakers when the topic is familiar
Use cognates and visual cues to derive meaning from texts that contain unfamiliar words, expressions and structures.

Read simple materials independently but have to guess at meanings of longer more complex material.

I used the Webb Alignment Model to conduct the final analysis of the coded data. I analyzed the alignment data under the categories (a) categorical concurrence, (b) depth-of-knowledge consistency, and (c) range-of-knowledge correspondence. When all panelists did not arrive at the same coding, the discussion and results were recorded in notes.

**Categorical Concurrence**

Categorical concurrence is a minimum requirement in alignment research. I examined the Checkpoint B syllabus’ broad content areas such as house and home. The total number of item-objective matches and the hits within a standard were averaged across all panelists to determine the average number of items per content category. There are fifteen content categories on the NYS Spanish Syllabus. Webb (2002) suggested six hits or six items to achieve scores of sufficient reliability. However, with a raised cut score of 65%, there has to be between nine and ten hits.

**Depth of Knowledge**

This subcategory of consistency compared the complexity of knowledge expressed in specific objectives within the syllabus. Each test item was matched to a cognitive area such as recall, skill and/or concept, strategic thinking, and extended thinking. (Webb 1999). The test items should be at the cognitive level outlined in the state syllabus. The cut score for passing the LOTE Spanish Comprehensive Regents is
65%, making 65% the total number of the number of items on the assessment that must target the skills as outlined by NYSED in the syllabus.

**Range of Knowledge**

I compared the breadth of the Checkpoint B syllabus with the breadth of the assessment by coding the number of objectives within the syllabus and matching them to assessment items. According to Webb (2002), to have sufficient alignment relative to range of knowledge, at least 50% of the objectives within a syllabus needed to be measured by at least one assessment item. Webb based the 50% match on the 50% score most students need to pass. However, for the Spanish LOTE examination, 65% is the number needed to pass: therefore, the matched number of objectives to one item on the assessment is 65% in this study.

**Reliability**

**Webb Alignment Criteria of Reliability of Coding**

I used the Webb’s Alignment Method to determine the alignment between a set of content objectives. The participants entered three kinds of data: depth-of-knowledge (DOK) levels of each content objective, DOK levels of each assessment item, and the standards/objective(s) targeted by each assessment item.

“Among the kinds of reliability--stability, reproducibility, and accuracy--reproducibility arguably is the strongest and most feasible kind to test. It amounts to evaluating whether a coding instrument, serving as common instructions to different observers of the same set of phenomena, yields the same data within a tolerable margin of error” (Krippendorff, 2004a, p. 78). Hayes (2005) posits that the more observers agree on the data they generate, and the larger the sample of units they describe, the more likely
it is that the data can be reproducible by an independent group of coders and that the
 coding is reproducible and trustworthy. The key to reliability is the agreement among
coders. The set of units used in the reliability data is a random sample from the universe
of data whose reliability is in question. The read-behind method was used in determining
the DOK levels of the content objectives and item match to standards by a process of
consensus among the panelists. The process depends on two panelists consistently
applying the DOK levels to the assessment items and coding the items to the New York
State Standards. Given such reliability data, a good index of reliability should have the
following properties:

1. It should assess the agreement between two observers who describe each of
the units of analysis separately from each other and be followed by a
discussion of the findings.

2. It should be grounded in the distribution of the categories actually used by the
observers.

3. Sampling behavior should be known and reported. A random sample is used
by the coders to conduct a reliability check. Random items within each
administration of the Spanish Comprehensive Regents Examination were
selected and coded. The coding results were compared to the original
findings (Webb, 2005).

Inter-rater reliability is at the heart of this method. The percentage of agreement is
calculated and disclosed in Chapter 4.
Summary

The Common Core State Standards coding process utilized two raters using a double rater read-behind process that allowed ongoing consensus discussion to occur (Sato, Lagunoff, & Worth, 2011). Webb's methodology (1999) was appealing because it offered both a flexible and comprehensive model that goes beyond simplistic content matching. Webb's procedures allow for averaging of ratings from individual panel members or using consensus.

Categorical concurrence shows whether the same content areas are covered by both the test and the academic standards. Depth of Knowledge is a scale with four levels ranging from simple recite/recall of facts (Level 1) to higher order thinking (Level 4).

The revised NYS Spanish Comprehensive Regents Examination was implemented by NYSED from 2003-2011. I chose the years 2006-2011 as a sample. I selected the Webb alignment method and two coders using a read-behind method to conduct this qualitative case study and arrive at findings.
CHAPTER 4

FINDINGS

Alignment between content standards and syllabus and standardized tests is a significant issue in test validation and pedagogy. The four alignment criteria are as follows: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation. Balance of knowledge representation is beyond the scope of this study. The purposes for this alignment study of the years 2006-2011, of Part 2 of the Spanish Languages Other Than English Comprehensive Regents Examination are to describe (a) the categorical concurrence of the New York State Comprehensive Spanish Regents Examination 2006-2011. Categorical concurrence refers to the proportion of overlap between the content stated in the standards document and the items assessed by the New York State Comprehensive Spanish Regents Examination; (b) the depth of knowledge consistency of the New York State Comprehensive Spanish Regents Examination 2006-2011, which is, the level of cognitive complexity, as defined by Webb’s Depth of Knowledge of the items of the New York State Comprehensive Spanish Regents Examination; (c) the range of knowledge correspondence of the New York State Comprehensive Spanish Regents Examination 2006-2011, defined as the number of content objectives of the New York State Checkpoint B syllabus assessed by items of the New York State Comprehensive Spanish Regents Examination.

“An alignment review can provide one form of evidence supporting the validity of a state’s assessment system. Alignment results should demonstrate that the assessment represent the full range of the content standards and the assessments measure student knowledge in the same manner and at the same level of
[cognitive] complexity as specified in the content standards.” (Taylor et al., 2009 p.1).

**Analysis Strategy**

I analyzed the data generated from coding the items to the New York State Modern Language Standards, New York State mandated syllabus and Webb’s Depth of Knowledge. I used the Webb Alignment Tool of coding the cognitive complexity of each of the items. Items from 2006 January and June administration of the New York State Comprehensive Spanish Regents Examination are coded to these three overarching concepts –Modern Language Standards, Syllabus and Webb’s depth of Knowledge. The findings were documented in an excel spreadsheet under the headings, Year/mth/day, Question number, Depth of Knowledge level, Depth of knowledge description, Depth of knowledge notes, Category number, Category/theme, Topic, Notes, Checkpoint, Standard, Notes.

“Throughout the act of scientific inquiry, concepts play a central role. They are significant elements in the prior scheme that the scholar has of the empirical world; they are likely to be the terms in which his problem is cast; they are the categories for which data are sought and in which data are grouped; they usually become the chief means for establishing relations between data; and they are the anchor points in interpretation of findings. (Blumer, 1969, p. 26, as cited in Corbin, Strauss, & Anselm, 2007, p. 51).

The state syllabus is divided into fifteen concepts or themes under which vocabulary and activities are linked. These concepts are: personal identification, house and home, services, family life, physical environment, meal taking/food/drink, health and
welfare, education, earning a living, public and private services, shopping, travel and current events.

Webb’s Depth of knowledge is divided into four concepts, recall, basic application, Strategic thinking, and extended thinking (Webb 2005). Using a read-behind method and through consistent coding of items to these concepts, I arrived at findings for the research questions.

Data Collection Methods and Data Analysis Construct

I applied the Webb Alignment Method to collect and analyze the alignment data. This alignment method has undergone substantial research (e.g., Webb, 1997; 1999; 2005), and it has been used successfully in approximately 15 other states and in two other countries. The results of the analyses are reported at the domain level, meaning how well the test items align with each of these broad content categories (Webb, 2005). The Webb method includes four major criteria to evaluate alignment. These criteria link with statistical procedures used to assess how well Part 2 of the assessments, syllabus and standards documents match. The four alignment criteria are as follows: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation. Balance of knowledge representation is beyond the scope of this study. The raters used a read behind method during the coding process.

Categorical Concurrence

Categorical concurrence is a basic measure of alignment between content standards and test items. This term refers to the proportion of overlap between the content stated in the standards document, the syllabus and that assessed by items on the test. There are fifteen content categories on the NYS Spanish Syllabus. Webb (2002)
suggested six hits or six items to achieve scores of sufficient reliability. However with a raised cut score of 65% there has to be between nine and ten hits.

**Depth-of-Knowledge Consistency**

The purpose of using depth of knowledge as a measure of alignment is to determine whether a test item and corresponding syllabus checkpoint are both written at the same level of cognitive complexity. The raters made two separate judgments about each item, one about cognitive complexity and one about the Checkpoint B syllabus. These two judgments are compared to determine whether the item is written at the same checkpoint level and cognitive complexity to which it is linked. Webb refers to his comparison as depth-of-knowledge consistency. Results were summarized in terms of the percentage of items with cognitive complexity ratings at or above (more complex than) the rating of the selected content objective. Webb suggests that at least 50% of the items should have complexity ratings at or above the level of the corresponding content objective. This 50% is based on a cut score of 50%. However, since the cut score for this examination is 65%, then the complexity match ratings need to be 65% for there to be depth of knowledge consistency.

**Range-of-Knowledge Correspondence**

Range of knowledge indicates the number of content objectives assessed by items. Webb’s minimum level of acceptability is a 50% match. This means that at least 50% of the objectives of Checkpoint B must be matched to at least 50% of assessment items. The percentage match is based on the cut score. Since the cut score is 65%, a match of 65% would be considered a minimum level of acceptability.
Statistical Methods

There are 33 items in Part 2 in each administration of the Comprehensive Spanish Regents Examination. Three items in Part 4 are designed at Level 4 depth of knowledge for all examinations, and the grading sample and rubric are designed to be at Checkpoint B.

There are 33 items in Part 2 of the New York State Comprehensive Examination in Spanish; to calculate the percentage of categorical concurrence, I added the number of items written at the Checkpoint B level. The total number of items written at the Checkpoint B level was divided by 33 and then multiplied by 100 to arrive at the percentage.

There are 33 items in Part 2 of the New York State Comprehensive Examination in Spanish; to calculate the percentage of depth-of-knowledge consistency, I added all items written at Level 2, Level 3, and Level 4 of Webb’s Depth of Knowledge. The total number of items written at Webb’s Depth of Knowledge, 2 though 4, were divided by 33, then multiplied by 100 to arrive at the percentage. Questions 31-33 are classified at the Level 2 depth of knowledge. Questions 31-33 require creation of a paragraph based on a stimulus. Webb defines creation and application of concepts as depth of knowledge Level 4; however, the level of vocabulary and requirements are part of the Checkpoint B requirements, as per NYSED specifications.

To calculate the range-of-knowledge correspondence, I matched each item to one of the 15 categories or domains in the New York State Modern Languages Syllabus. Items were grouped in the corresponding content domains. Items tapping the same domain were classified as one match. I added the number of domains with match items
and divided by 15. This number was multiplied by 100 to arrive at the percentage of range-of-knowledge correspondence.

Reliability Data

The read-behind method was used to generate the data for this qualitative study on content validity and alignment of the New York State Spanish Comprehensive Regents Examination 2006-2011. To determine the reliability of the coding, two items were selected at random and each coder rated the items separately. The panel rated 24 items for the reliability check.

At the end of the rating session, the panelists discussed findings. I compared the 24 items coded in the reliability check to the items previously rated by the panelists. For the inter-rater check, panelists had identical coding 21 times from a total of 24 items coded. That means there is an 87.5% chance that coders will agree without having to come to consensus on the coding of items.

The coding of the 24 items was then compared to the coding conducted previously. I calculated the percentage of agreement to determine inter-rater reliability. Table 4 shows the percentage of inter-rater agreement in the following domains: Depth of Knowledge, Category/Content Domain (range of knowledge), Checkpoint, and Standard (categorical concurrence) as compared to the original coding for this study.

Below is the data reliability set which shows the percentage of agreement. The data reliability set below displays the percentage of rater agreement when comparing the original coding of items to the 24-item data set used to establish inter-rater reliability for Part 2 of the NYS Spanish Comprehensive Regents, 2006-2011.
Table 4

Data Reliability Set

<table>
<thead>
<tr>
<th>Depth of knowledge</th>
<th>Category/Content Domain</th>
<th>Checkpoint</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>83.3%</td>
<td>95.8%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5 displays the inter-rater reliability findings for Depth of Knowledge and Category/Content Domain. Checkpoint and Standard had 100% inter-rater agreement with original coding as well as the 100% agreement with items selected for the reliability check.

Table 5

Data Reliability Set for Depth-of-Knowledge and Category/Content Domains

<table>
<thead>
<tr>
<th>Year</th>
<th>Item Number</th>
<th>Rater Agreement With Original coding</th>
<th>Depth of Knowledge Rater Agreement</th>
<th>Rater Agreement With Original coding</th>
<th>Category/Content Domain Rater Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 2006</td>
<td>2</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
<td>Rater 1 No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
<td>Rater 2 No</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
<td>Rater 1 no</td>
<td>Rater 1 yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
<td>Rater 2 no</td>
<td>Rater 2 yes</td>
</tr>
<tr>
<td>June 2006</td>
<td>20</td>
<td>Rater 1 No</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rater 2 No</td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
</tr>
<tr>
<td>Jan 2007</td>
<td>10</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
<td>Rater 1 yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
<td>Rater 2 yes</td>
</tr>
<tr>
<td>Date</td>
<td>Rater 1</td>
<td>Rater 2</td>
<td>Rater 1</td>
<td>Rater 2</td>
<td>Rater 1</td>
</tr>
<tr>
<td>------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>Yes</td>
</tr>
<tr>
<td>June 2007</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>June 2007</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 2008</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 2008</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>June 2008</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 2009</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 2009</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 2009</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>Rater 1</td>
<td>Rater 2</td>
<td>Rater 1</td>
<td>Rater 2</td>
<td>Rater 1</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>June 2009</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 2010</td>
<td>No</td>
<td>No</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>June 2010</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 2011</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>June 2011</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Inter-coder reliability, also called inter-rater agreement is at the heart of the data reliability verification. Webb (2002) suggested six hits or six items per standard/domain to achieve scores of sufficient reliability. The percentage of inter-rater reliability is at acceptable limits to arrive at valid findings for this study.

**Data Analysis Summary**

Bachman and Palmer (2002) treat language use as a complex and highly interactive activity involving the “negotiation of intended meanings between two or more individuals in a particular situation” (p. 34). They define language ability as a “capacity that enables language users to create and interpret discourse (p. 33) and suggest that it includes two components, language knowledge and strategic competence; but they also recognize a need to consider “personal attributes, topical knowledge, affective schemata, and cognitive strategies (p. 33).

The New York State Modern Language Standards are very broad learning goals. Items are aligned to Standard 1 in the skills of listening and speaking and reading and writing.

Standard 1: Students will be able to use a language other than English for communication.

**Checkpoint B - Listening and Speaking**

- Comprehend messages and short conversations when listening to peers, familiar adults, and providers of public services either in face-to-face interactions or on the telephone
• Understand the main idea and some discrete information in television, radio, or live presentations

• Initiate and sustain conversations, face to face or on the phone, with native-Speaking or more fluent individuals

• Select vocabulary appropriate to a range of topics, employ simple and complex sentences in present, past, and future time frames, and express details and nuances by using appropriate modifiers

• Exhibit spontaneity in their interactions, particularly when the topic is familiar, but often rely on familiar utterances

• Use repetition and circumlocution as well as gestures and other nonverbal cues to sustain conversation

Checkpoint B - Reading and Writing

• Read and comprehend materials written for native speakers when the topic and language are familiar

• Use cognates and contextual and visual cues to derive meaning from texts that contain unfamiliar words, expressions, and structures

• Read simple materials independently, but may have to guess at meanings of longer or more complex material

• Write short notes, uncomplicated personal and business letters, brief journals, and short reports

• Write brief analyses of more complex content when given the opportunity for organization and advance preparation, though errors may occur more frequently
- Produce written narratives and expressions of opinion about radio and
television programs, newspaper and magazine articles, and selected stories,
songs, and literature of the target language.

Standard 2 is not explicitly tested in the items in Part 2 of the New York State Spanish
Comprehensive Regents Examination.

Checkpoint B - Cross Cultural Skills and Communication
- Exhibit more comprehensive knowledge of cultural traits and patterns
- Draw comparisons between societies
- Recognize that there are important linguistic and cultural variations among
groups that speak the same target language
- Understand how words, body language, rituals, and social interactions
  influence communication.

“There is a widespread popular belief that any person who takes a test either
passes or fails it . . . This is patently false . . . A second popular belief is that when a test
is used to pass or fail examinees, the distinction between the two outcomes is clear-cut
and unequivocal. This is almost never true (Ebel, 1979, p. 337). Scores on most tests
reveal differences of degree. Choosing a cut score does not transform a test into an
instrument that detects differences in kind. A person with a score just above the cut is not
very different from one with a score just below the cut--until someone makes them
different by treating them differently. Because tests are not perfectly reliable and valid,
whenever a cut score is set, some of the resulting decisions will not be correct (Dwyer,
1996, p. 361). The New York State Education Department raised the cut score for
students entering high school in 2002 from 55% to 65%; however, the test items were not
realigned to correspond to the new cut scores. The June 2008 administration of the Spanish Comprehensive Regents Examination is the only administration of the years 2006-2011 that was constructed with full alignment.

Findings

Findings from the analysis of the alignment of the items to the content syllabus state standards and depth of knowledge conducted in this study provide evidence of varying quality control and content validity of the New York State Spanish Comprehensive Regents Examination.

The current study provides an important source of content validity and alignment studies as well as the utility of the Webb alignment tool in standard setting judgments.

Table 6 displays a summary of alignment outcomes based on Webb’s criterion for categorical concurrence, depth-of-knowledge consistency and range-of-knowledge consistency for Part 2 of the New York State Comprehensive Spanish Regents Examinations 2006-2011. The range for categorical concurrence was 51.51% to 90.09%. The range of depth-of-knowledge consistency was 48.48% to 72.27% and the range of the range-of-knowledge correspondence was 46.6% to 80%.

Table 6

*Summaries of Alignment Outcomes on Each of Webb’s Criterion, NYS Comprehensive Spanish Regents Examination 2006-2011*

<table>
<thead>
<tr>
<th>Year</th>
<th>Categorical Concurrence</th>
<th>Depth of Knowledge Consistency</th>
<th>Range of Knowledge Correspondence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 2006</td>
<td>51.51%</td>
<td>48.48%</td>
<td>80%</td>
</tr>
<tr>
<td>June 2006</td>
<td>81.81%</td>
<td>51.51%</td>
<td>66.66%</td>
</tr>
</tbody>
</table>
According to the New York State Report Card, the passing percentage or the percentage of students scoring 65% and over were as follows:

<table>
<thead>
<tr>
<th>Year</th>
<th>January</th>
<th>June</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>60.60%</td>
<td>51.51%</td>
</tr>
<tr>
<td>2007</td>
<td>51.51%</td>
<td>54.54%</td>
</tr>
<tr>
<td>2008</td>
<td>75.75%</td>
<td>60.60%</td>
</tr>
<tr>
<td>2009</td>
<td>72.72%</td>
<td>72.72%</td>
</tr>
<tr>
<td>2010</td>
<td>81.81%</td>
<td>60.60%</td>
</tr>
<tr>
<td>2011</td>
<td>90.90%</td>
<td>51.51%</td>
</tr>
</tbody>
</table>

In reporting the passing scores, both January and June administration scores were combined. The percentage passing for each administration was combined and divided by the number of administrations for the year. This does not permit an analysis of scores by individual administrations, as the state scores are reported once per calendar year with the
averages of the percentages for each administration combined. The alignment findings fluctuated without any patterns for each administration of the New York State Spanish Comprehensive Regents Examination for the years 2006-2011.

**Research Question 1**

At what level of cognitive function as defined by Webb’s Depth of Knowledge are the performance tasks for the LOTE Spanish Comprehensive Regents Examination?

The three-point scale used to describe cognitive alignment (Depth-of-Knowledge consistency-DOK), as defined by Webb, ranges from full to insufficient and is defined as follows:

- **Full** - For each major topic, the most cognitively demanding expected performance for all students is comparable to the most cognitively demanding assessment activity taken by all students.

- **Acceptable** - For nearly all major topics, the most cognitively demanding expected performance for all students is comparable to or can be inferred from the most cognitively demanding assessment activity taken by all students.

- **Insufficient** - Students can be judged as performing at an acceptable level on the assessments without having to demonstrate for any topic the attainment of the most cognitively demanding expected performance for all students.

NYSED instituted a cut score of 65% as the passing rate, which under Webb’s criteria means that full alignment takes pace at 65% and higher. The June 2008 administration is the only year of the assessments under study (2006-2011) to contain full alignment. The depth-of-knowledge consistency for the years 2006-2011 varied in range.
from 51.51% to 72.72%, with June 2008 representing the only administration of the period under review to have over 65% depth-of-knowledge consistency. The June 2008 administration had a depth-of-knowledge consistency of 72.72%.

Table 8 displays the depth-of-knowledge consistency for the NYS Spanish Comprehensive Regents Examination for the period under study, 2006-2011.

<table>
<thead>
<tr>
<th>Year</th>
<th>Depth of Knowledge Consistency</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 2006</td>
<td>48.48%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2006</td>
<td>51.51%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Jan. 2007</td>
<td>51.51%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2007</td>
<td>54.54%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Jan. 2008</td>
<td>60.60%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2008</td>
<td>72.72%</td>
<td>Full</td>
</tr>
<tr>
<td>Jan. 2009</td>
<td>60.60%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2009</td>
<td>51.51%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Jan. 2010</td>
<td>48.48%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2010</td>
<td>60.60%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Jan. 2011</td>
<td>54.54%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2011</td>
<td>60.60%</td>
<td>Insufficient</td>
</tr>
</tbody>
</table>

**Research Question 2**

How aligned are the test questions to the mandated syllabus, Checkpoint B, for the years 2006-2011?

The three-point scale used to describe Range-of-Knowledge Correspondence ranges from full to insufficient and is defined below:

- Full - The percentage of hit is above the cut score.
- Acceptable - The percentage of hits is within the range of the cut scores, 60%-
65%.

- Insufficient - The percentage of hits shows a narrow range of curricular items assessed (below NYSED’s cut score of 65%).

NYSED instituted a cut score of 65% as the passing rate, which under Webb’s criteria means that full alignment takes place at 65% and higher. Student grades between the ranges of 60%-64% were eligible for appeal. This makes the acceptable range of alignment between 60% and 64%. The Range of Knowledge Correspondence for the period under study, 2006-2011 ranges from 46.66% to 80%. January 2011 is the only administration for the period under review that contains insufficient alignment.

Table 9 displays the Range-of-Knowledge correspondence for Part 2 of the NYS Comprehensive Spanish Examination for the years 2006-2011.

Table 9

*Range-of-Knowledge Correspondence, Part 2 of The New York State Comprehensive Spanish Regents Examinations 2006-2011*

<table>
<thead>
<tr>
<th>Year</th>
<th>Range of Knowledge Correspondence Items 1-30</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 2006</td>
<td>80%</td>
<td>Full</td>
</tr>
<tr>
<td>June 2006</td>
<td>66.66%</td>
<td>Full</td>
</tr>
<tr>
<td>Jan. 2007</td>
<td>60%</td>
<td>Acceptable</td>
</tr>
<tr>
<td>June 2007</td>
<td>73.3%</td>
<td>Full</td>
</tr>
<tr>
<td>Jan. 2008</td>
<td>60%</td>
<td>Acceptable</td>
</tr>
<tr>
<td>June 2008</td>
<td>73.3%</td>
<td>Full</td>
</tr>
<tr>
<td>Jan. 2009</td>
<td>66.66%</td>
<td>Full</td>
</tr>
<tr>
<td>June 2009</td>
<td>80%</td>
<td>Full</td>
</tr>
<tr>
<td>Jan. 2010</td>
<td>66.66%</td>
<td>Full</td>
</tr>
<tr>
<td>June 2010</td>
<td>60%</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Jan. 2011</td>
<td>46.66%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2011</td>
<td>66.66%</td>
<td>Full</td>
</tr>
</tbody>
</table>
Research Question 3

What percentage of the mandated syllabus is assessed by Part 2 of the LOTE Spanish Comprehensive Regents Examination from the years 2006-2011?

The total number of item-objective matches and the hits within a syllabus category were averaged across all panelists to determine the average number of items per category.

There are fifteen content categories on the NYS Spanish syllabus. NYSED instituted a cut score of 65% as the passing rate, which under Webb’s criteria means that full alignment takes place at 65% and higher. Grades between the ranges of 60%-64% were eligible for appeal. This makes the acceptable range of alignment between 60% and 64%. The ranges of Categorical concurrence for the period under review, 2006-2011, were from 51.51% to 90.90%. The June 2007 administration is the only year for the period under review to contain insufficient alignment in the domain of categorical concurrence.

Table 10 displays the categorical concurrence of the NYS Comprehensive Spanish Regents Examination for the years 2006-2011.

Table 10

<table>
<thead>
<tr>
<th>Year</th>
<th>Categorical Concurrence</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 2006</td>
<td>51.51%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2006</td>
<td>81.81%</td>
<td>Full</td>
</tr>
<tr>
<td>Jan. 2007</td>
<td>60.60%</td>
<td>Acceptable</td>
</tr>
<tr>
<td>June 2007</td>
<td>51.51%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Jan. 2008</td>
<td>75.75%</td>
<td>Full</td>
</tr>
<tr>
<td>June 2008</td>
<td>72.72%</td>
<td>Full</td>
</tr>
<tr>
<td>Jan. 2009</td>
<td>81.81%</td>
<td>Full</td>
</tr>
<tr>
<td>June 2009</td>
<td>90.90%</td>
<td>Full</td>
</tr>
<tr>
<td>Jan. 2010</td>
<td>84.84%</td>
<td>Full</td>
</tr>
</tbody>
</table>
Alignment Findings

Displayed in Table 11 is a summary of the alignment findings. The findings are displayed under Webb’s Alignment Criteria of Categorical Concurrence, Depth of Knowledge Consistency and Range of Knowledge Correspondence. Balance of Knowledge Representation is beyond the score of this study. The cut score for the NYS Spanish Comprehensive Regents Examination is 65%. There the percentages in each of Webb’s alignment categories must be 65% or higher in order to arrive at alignment. Grades of 60% can be appealed during this test administration period, and consequently 60% is recorded as acceptable alignment.

Table 11 presents the findings on alignment outcomes for Part 2 of the New York State comprehensive Spanish Regents January and June examinations for the years 2006 to 2011. The alignment outcomes are based on Webb’s Criteria for Categorical Concurrence, Depth-of-Knowledge Consistency and Range-of-Knowledge Consistency. The categorical concurrence was highest in June 2009 at 90.9% and lowest in January 2006 at 51.51%. The worst year was 2007, with recorded categorical concurrence of 60.60% and 51.51% in January and June, respectively. The best year was 2009, with recorded categorical concurrence of 81.81% and 90.90% in January and June, respectively. The categorical concurrence between January 2009 and June 2011 were the highest compared with the period January 2006 and June 2008.

The January 2006 administration also scored the lowest for depth-of-knowledge, 48.48%, which was the same percentage scored at the January 2010 exam. However,
although categorical concurrence and depth of knowledge in January 2006 were the lowest in the period under consideration, the range-of-knowledge correspondence was the highest overall at 80%. Indeed, the range-of-knowledge correspondence was 80% at the June 2009 exam, which corresponded with a 51.51% depth-of-knowledge consistency and the 90.90% categorical concurrence in the same year. The range-of-knowledge correspondence recorded its lowest level in January 2011 at 46.66%, which corresponded with a 54.54% depth-of-knowledge consistency, and 87.87% categorical concurrence.

The depth-of-knowledge consistency was aligned in full only in June 2008, at 72.72%, at the same exam at which the categorical concurrence was 72.72% and the range-of-knowledge correspondence was 73.30% (see Table 5). In every other case, the alignment was insufficient. Thus, the depth-of-knowledge consistency was at its highest point when alignment was full, but it corresponded to the third lowest percentage of categorical concurrence.

The situation was different with regard to the range-of-knowledge correspondence in the same period (see Table 9). In this case, alignment was full at both exams in 2006 and 2009, and also in June 2007, June 2008, January 2010, and June 2011. It was acceptable in January 2007, January 2008, and June 2010. The June exams have recorded fuller alignment compared with the January exam, five compared to two. Of the three exams whose alignment were acceptable, two were held in January and one in June. The only exam whose alignment was insufficient was held in January. Closer attention needs to be placed on the January exams to understand why they have a lower alignment level than the June exams.
The full alignment of range-of-knowledge correspondence of 80% in 2006 corresponded to the lowest levels of categorical concurrence and depth-of-knowledge consistency in the period under review.

Table 11 displays the alignment of Part 2 of the NYS Comprehensive Spanish Regents Examination for the years 2006-2011. It is a summative display of the Categorical Concurrence, Depth-of-Knowledge Consistency, and Range-of-Knowledge Correspondence.

**Table 11**

*Alignment of Part 2 of the New York State Comprehensive Spanish Regents Examinations 2006-2011*

<table>
<thead>
<tr>
<th>Year</th>
<th>Categorical Concurrence</th>
<th>Depth of Knowledge Consistency</th>
<th>Range of Knowledge Correspondence Items 1-30</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 2006</td>
<td>51.51%</td>
<td>48.48%</td>
<td>80%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2006</td>
<td>81.81%</td>
<td>51.51%</td>
<td>66.66%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Jan. 2007</td>
<td>60.60%</td>
<td>51.51%</td>
<td>60%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2007</td>
<td>51.51%</td>
<td>54.54%</td>
<td>73.3%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Jan. 2008</td>
<td>75.75%</td>
<td>60.60%</td>
<td>60%</td>
<td>Acceptable</td>
</tr>
<tr>
<td>June 2008</td>
<td>72.72%</td>
<td>72.72%</td>
<td>73.3%</td>
<td>Full</td>
</tr>
<tr>
<td>Jan. 2009</td>
<td>81.81%</td>
<td>60.60%</td>
<td>66.66%</td>
<td>Acceptable</td>
</tr>
<tr>
<td>June 2009</td>
<td>90.90%</td>
<td>51.51%</td>
<td>80%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Jan. 2010</td>
<td>84.84%</td>
<td>48.48%</td>
<td>66.66%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2010</td>
<td>66.66%</td>
<td>60.60%</td>
<td>60%</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Jan. 2011</td>
<td>87.87%</td>
<td>54.54%</td>
<td>46.66%</td>
<td>Insufficient</td>
</tr>
<tr>
<td>June 2011</td>
<td>81.81%</td>
<td>60.60%</td>
<td>66.66%</td>
<td>Acceptable</td>
</tr>
</tbody>
</table>

**Content Validity Findings**

Content validity generally refers to the degree to which a test approximately represents the content domain it is intended to measure (Martone & Sireci, 2009). An academic achievement test is considered content valid if and when (a) the curriculum universe has been defined (called the "content domain") and (b) the test adequately
samples that universe. Using this definition of content validity, the Spanish
Comprehensive Regents Examinations 2006-2011 is content valid with the exception of
the administrations in January, 2006 and June 2007, when the categorical concurrence or
content domain falls below the range 60%. (cut score 65%, appeal score 60%).

Table 12 displays the content validity Part 2 of the New York State Spanish
Comprehensive regents Examination for the years 2006-2011. The areas of Categorical
Concurrence and Range-of-Knowledge Correspondence are displayed. NYSED
designated 65% as the cut score for passing this examination. Part 2 of the Spanish
Comprehensive Regents Examinations 2006-2011 is content valid with the exception of
the administrations in January 2006 and June 2007.

Table 12

<table>
<thead>
<tr>
<th>Year</th>
<th>Categorical Concurrence</th>
<th>Range of Knowledge Correspondence Items 1-30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 2006</td>
<td>51.51%</td>
<td>80%</td>
</tr>
<tr>
<td>June 2006</td>
<td>81.81%</td>
<td>66.66%</td>
</tr>
<tr>
<td>Jan. 2007</td>
<td>60.60%</td>
<td>60%</td>
</tr>
<tr>
<td>June 2007</td>
<td>51.51%</td>
<td>73.3%</td>
</tr>
<tr>
<td>Jan. 2008</td>
<td>75.75%</td>
<td>60%</td>
</tr>
<tr>
<td>June 2008</td>
<td>72.72%</td>
<td>73.3%</td>
</tr>
<tr>
<td>Jan. 2009</td>
<td>81.81%</td>
<td>66.66%</td>
</tr>
<tr>
<td>June 2009</td>
<td>90.90%</td>
<td>80%</td>
</tr>
<tr>
<td>Jan. 2010</td>
<td>84.84%</td>
<td>66.66%</td>
</tr>
<tr>
<td>June 2010</td>
<td>66.66%</td>
<td>60%</td>
</tr>
<tr>
<td>Jan. 2011</td>
<td>87.87%</td>
<td>46.66%</td>
</tr>
<tr>
<td>June 2011</td>
<td>81.81%</td>
<td>66.66%</td>
</tr>
</tbody>
</table>

Beyond the general definition of content validity, with respect to a content
validity study, there are at least four potential aspects: domain definition, domain
representation, domain relevance, and appropriateness of the test construction procedures (Sireci, 1998a, 1998b).

Domain definition is the process used to define operationally the content domain tested. This is derived from the state-established curriculum frameworks, which for Languages Other Than English is the checkpoint syllabus and the four skills outlined, which are speaking, listening, reading, and writing. The New York State Spanish Comprehensive Regents Examination has face validity with regard to the skills of listening, reading, and writing. Each of these skill areas is measured in a consistent manner for the period under review (2006-2011). The four skills of speaking, listening, reading, and writing are constructed in a consistent manner, with the points in each category awarded in a consistent manner. The construction of the test corresponds to the outline and skills specified by NYSED. Of over 300 items for the assessments under review (2006-2011), fewer than five (less than 1.6%) did not correspond to any of the categories and topics outlined in the syllabus. The domains are relevant to the skills students need to function in a career or academic environment, and to communicate when traveling or encountering a native speaker of the target language. I cannot comment on the appropriateness of test construction procedures, as those could not be observed due to security protocols in place by NYSED.

The domain representation refers to the degree to which the test represents and adequately measures all facets of the intended content domain. To evaluate the domain representation of the Comprehensive Spanish Regents Examination, I inspected the items on the test and analyzed the degree to which they were aligned with the test specification (Crocker, Miller, & Franks, 1989). January 2006 and June 2007 are the only
administrations with insufficient domain definitions for the period under study, the January and June administrations for the years 2006-2011.

Content validity can be supported if there were strong quality control measures in place that correspond to the domains. Content validity emerged in the field of testing to guard against strictly numerical evaluations of tests that overlooked serious threats to the validity of inferences derived from test scores (Sireci, 1998).

Polikoff, Porter, and Smithson (2011) posit, “State content standards and corresponding assessments of student achievement are the foundations upon which the current system of standards-based accountability in U.S. education is built. In the initial formulation of standards-based reform, as well as in the No Child Left Behind law that codified standards-based reform in federal law, the coherence of standards and assessments was seen as paramount to ensuring the validity of the interpretations made from student test scores. There have been a number of alignment procedures developed to measure the extent to which the standards and assessments are, in fact, coherent” (p. 25). This study used the Webb Alignment Method. One of the goals of testing is to improve instruction. However, if we are to assume that testing drives instruction, then the misalignment of tests does not accurately measure what ought to be measured, the syllabus. This raises concerns about the uses of the test and the consequences associated with results from the assessment. The level of cognitive rigor is consistently below the cut score determined by NYSED (65%) with the exception of the June 2008 administration as displayed in Table 13.
Table 13

*NYS Spanish Comprehensive Regents Examination, Cognitive Rigor 2006-2011*

<table>
<thead>
<tr>
<th>Date of Administration of NYS Comprehensive Regents</th>
<th>Depth of Knowledge Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 2006</td>
<td>48.48%</td>
</tr>
<tr>
<td>June 2006</td>
<td>51.51%</td>
</tr>
<tr>
<td>Jan. 2007</td>
<td>51.51%</td>
</tr>
<tr>
<td>June 2007</td>
<td>54.54%</td>
</tr>
<tr>
<td>Jan. 2008</td>
<td>60.60%</td>
</tr>
<tr>
<td>June 2008</td>
<td>72.72%</td>
</tr>
<tr>
<td>Jan. 2009</td>
<td>60.60%</td>
</tr>
<tr>
<td>June 2009</td>
<td>51.51%</td>
</tr>
<tr>
<td>Jan. 2010</td>
<td>48.48%</td>
</tr>
<tr>
<td>June 2010</td>
<td>60.60%</td>
</tr>
<tr>
<td>Jan. 2011</td>
<td>54.54%</td>
</tr>
<tr>
<td>June 2011</td>
<td>60.60%</td>
</tr>
</tbody>
</table>

**Summary**

Content validity is the correspondence between curriculum objectives and the objectives being assessed. To put it another way, it is an aspect of construct validity that emphasizes evidence bearing on the appropriateness of the knowledge, skills, and abilities measured by an assessment. Content validity is comprised of (1) the sampling adequacy of test content, (2) the sampling adequacy of test responses, (3) the relevance of test content to a content universe, (4) the relevance of test responses to a behavioral universe, (5) the clarity of content domain definitions, and (6) the technical quality of test items.

There are two major characteristics of alignment: topical and conceptual and cognitive complexity, or demand emphasis. The topical and conceptual or the New York State Checkpoint B syllabus specifies what students are required to know. The cognitive
complexity, or demand emphasis as taxonomized by Webb’s Depth of Knowledge, indicates the cognitive process required to arrive at a solution to the stimuli.

The cognitive rigor of the NYS Spanish Comprehensive Regents Examination 2006-2011 is an area of concern in future test development cycles. The alignment of items to the syllabus and state standards ought to be more consistent so that decisions and judgments made on the basis of test scores are related to a body of knowledge, mandated standards and cut scores. Validating tests is an ongoing task and when stakes are high for students, personnel, and school communities, it is important that alignment is not left to chance but rather is a part of the process of test development.

Chapter 5 includes a discussion of the key findings, conclusions, implications for policy and practice, and recommendations for further study.
CHAPTER 5

DISCUSSION, RECOMMENDATIONS, AND CONCLUSIONS

Discussion

The purposes for this alignment study of the years 2006-2011 of Part 2 of the Spanish Languages Other Than English Comprehensive Regents Examination are to describe (a) the categorical concurrence of the New York State Comprehensive Spanish Regents Examination 2006-2011, which refers to the proportion of overlap between the content stated in the standards document and the items assessed by the New York State Comprehensive Spanish Regents Examination; (b) the depth-of-knowledge consistency of the New York State Comprehensive Spanish Regents Examination 2006-2011, which is the level of cognitive complexity as defined by Webb’s Depth of Knowledge of the
items of the New York State Comprehensive Spanish Regents Examination; and (c) the range of knowledge correspondence of the New York State Comprehensive Spanish Regents Examination 2006-2011, defined as the number of content objectives of the New York State Checkpoint B syllabus assessed by items of the New York State Comprehensive Spanish Regents Examination.

The research questions for this qualitative study were as follows:

1. What is the categorical concurrence and proportion of the overlap of Part 2 of the New York State Comprehensive Spanish Regents Examination 2006-2011?

2. What is the depth-of-knowledge consistency as defined by Webb’s Depth of Knowledge of the New York State Comprehensive Spanish Regents Examination items for the years 2006-2011?

3. What is the range-of-knowledge correspondence, defined as the number of content objectives of the New York State Checkpoint B syllabus assessed by items in Part 2 of the New York State Comprehensive Spanish Regents Examination 2006-2011?

New York State Education Department personnel agreed to be a part of the Common Core State Standards and to receive the federal funding from Race to the Top. This requires adopting new standards, new teacher evaluations, and new assessments. Part of a teacher’s evaluation will be based on standardized test scores. New York State Education Department personnel plan to create new assessments aligned to the Common Core State Standards beginning in the 2013 school year. No one has seen samples of the new tests, and no psychometric data on the tests were released.

Forty-two states have signed on to implement the Common Core State Standards
and to implement testing created by one of the two consortia. Dunbar, Koretz, and Hoover (1991) noted, “The nation stands poised on the brink of yet another wave of test-based reform, and again we appear prepared to undertake it without sufficient quality control” (p. 302).

The results from this study indicate that simply creating a policy to raise the cut score does not in itself indicate more cognitive and academic rigor or increased quality of the assessment instrument as measured by the degree of alignment. Using assessments for accountability purposes only begins to be rational if the assessment measures the required learning as outlined in the syllabus. Quality control can be achieved by ensuring that the assessment measures what is supposed to be measured according to state statutes and laws and that the items are aligned to the correct cognitive level as determined by the NYSED policy. This study adds to the complement of alignment studies using the Webb Alignment Method. The findings of this case study are also consistent with studies conducted by Webb on large-scale assessments that show lower levels of cognitive demand on the assessments than indicated by the cut score. In addition, this qualitative case study provides data and recommendations on the content validity and alignment of Part 2 of the NYS Comprehensive Spanish Regents Examinations 2006-2011 and the alignment of the assessments to the standards and syllabus.

The results are consistent with findings from alignment studies conducted by Webb et al. (2005) and Porter (2009). Webb (1999) developed the criteria for alignment in conjunction with the Council of Chief State School Officers to analyze alignment as it applied to Mathematics and Science in four states (Webb, 1999). Subsequently, he
applied his alignment method to language arts standards and assessment in three states and to the Wisconsin standards in its alternative assessment for students with disabilities. Webb’s findings were that items on the state tests demanded a lower level of cognitive knowledge and skills than the standards expected (Rothman, 2003, p. 4). Findings from this study on the alignment of the NYS Spanish Comprehensive Regents Examination show that the cognitive demand is consistently below the expected score of 65%. In fact, June 2008 is the only administration with a cognitive demand definitively above the cut score. In Webb’s study, “In four of the fourteen tests analyzed, the degree of consistency was quite high; in two cases, a fourth grade mathematics test and a sixth grade mathematics test, the match was 100%. For the rest, however, the test items were rated at lower levels than the standards” (Rothman, 2003, p. 4).

Webb posited that the inadequate number of items assessing the higher levels of depth of knowledge was the major reason for insufficient alignment. Based on this observation, Webb (2006) recommended replacement of lower level depth-of-knowledge items for the assessment to reach acceptable levels of alignment. I recommend the replacement of lower depth-of-knowledge items with higher depth-of-knowledge items for the Spanish Comprehensive Regents Examination.

Webb also found that the tests and standards tended to lack range-of-knowledge correspondence. To meet this criterion, tests must include items that measure at least half of the related objectives within a given standard. However, the study found that test items tended to cluster around a few objectives, leaving most of the objectives un-assessed (Webb, 1999). The findings were similar in the studies of English language arts standards and tests. On the other hand, the Wisconsin Alternative Assessment for
students with disabilities was relatively well aligned with state standards (Rothman, 2003, p. 4). This study on the alignment of Part 2 of the NYS Spanish Comprehensive Regents Examination found that the items for the period under study, 2006-2011, meet the criterion of acceptable range-of-knowledge correspondence, with the exception of the January 2011 administration.

Issues of items at the correct depth of knowledge, the acceptable range of knowledge of items, or construct under representation undermine the construct validity of the assessment. The failure of the assessment to adequately sample the domains leads to a systematic failure of the construct. Items on the assessment that do not appear on the syllabus raise the question of the quality control of the item bank. Items in the item bank need to be reviewed for alignment to the syllabus and coded at the correct depth of knowledge and categorical concurrence as one of the early steps of test construction.

This study produced results that align with concepts from the literature. The LOTE Spanish Comprehensive Regents Examinations over time are indicative of a narrowing of the range of knowledge of the syllabus in the January 2011 administration and in Part 4, the essay portion of the assessment, with each successive administration of the examination. The Part 4 essay can be answered with Checkpoint A vocabulary as the years progress. The use of items for Part 4 that can be answered with Checkpoint A vocabulary enables students to gain a passing score without writing at the correct level of difficulty and complexity. The decision to accept or reject the checkpoint level is at the discretion of the classroom teacher who teaches the class and scores the assessment.

This study is also consistent with the literature on large-scale assessments in that the details of the setting of cut scores, scoring errors, and measurements are not disclosed.
by state education departments for each large-scale assessment administered, which calls into question the construct validity of the assessment. Messick (1990) posited that construct validity of score meaning is the integrating force that unifies validity issues into a unitary concept. Assuming that validity of the assessment is at stake, the process of establishing cut scores and the standard error of measurement must be disclosed and transparent. State education departments use standardized testing as one of the currencies of reform. However, “Deep structures and practices must be transformed, not reformed” (Deal, 1990); and whether or not high stakes tests can lead to improvement in teaching and learning is still a matter of dispute.

**Recommendations for Policy**

Test scores or quantitative performance indicators exist in relation to the mandated standards and syllabus. There must exist conceptual links between a test taker’s performance on an assessment and other variables such as “a qualitative description or quantitative score, an interpretation of the ability under assessment, the decisions that will be made, and the consequences of the decisions that are made” (Bachman & Palmer, p. 30). Test construction is not a one-shot event. I argue that the process begins with adequate funding and realistic timelines to accommodate piloting and review.

**Cut Scores**

Kane (2010) posits that most high-stakes testing programs employ decision rules in which cut scores play a major role. Traditionally, the cut scores were determined by
analyzing the utilities (or consequences) associated with the adoption of different cut scores (Cronbach & Gleser, 1965), where the utilities were defined in terms of such outcomes as the productivity of workers hired or the performances of students placed into different courses. NYSED Commissioner and personnel instituted the Modern Language Standards in 1996, and the revised LOTE assessment was administered in 2003; at that time, the cut score was 50%. In subsequent years, NYSED personnel raised the cut score on the existing LOTE test to 65% as a sign of “get tough” policies for children and teachers. This should have involved an internal realignment of the domains to match the new cut score of 65%. This did not occur. Raising the cut score proved problematic because the alignment between the cut score and the new standards and syllabus was not verified. The new cut score existed on paper as part of education policy with consequences for students and practitioners. The assessment instrument needed to be psychometrically evaluated by panels of subject matter experts, psychometricians, and industrial psychologists to ensure alignment and congruence with the new policy.

Findings from this study show that simply raising the cut score does not imply raising the level of academic and cognitive rigor, and the act of raising the cut score of the assessment without first determining the level of alignment with the new standards is problematic for students and practitioners. The results from this study suggest that the alignment between the new standards, syllabus, cognitive rigor, and the cut scores are an important piece of information in the interpretation of the numerical construct of the test score.

High stakes testing or testing students for the purpose of accountability needs to be a transparent process. Not only must the assessment measure what it is intended to
measure, but the process of test construction must also be transparent. The reliability and technical report and the psychometric data of each assessment should be released to practitioners and researchers. If the public is told to measure schools and teachers and children’s learning on test scores, then the full breadth of what the test score means and implies should also be available. Systems of accountability do not only function downstream at the end product, but begin with quality control from the inception of the test construction process and with adequate funding and transparency.

**Standard Error Measurement**

The difference between test and retest for individuals can be caused by measurement error. The standard deviation of the measurements errors, true scores ± measurement error is the statistic known as the standard error of measurement. There is no publically released data on the stability and reliability of the LOTE assessment. NYSED personnel do not disclose the standard error measurement for any administration of the LOTE Spanish Comprehensive Regents Examination. Before 2010, NYSED permitted rescoring of individual papers without much oversight. By 2011 in New York City, no rescoring was permitted by the personnel in the Office of Assessment, without signatures of multiple raters. However, despite these new protocols that may minimize score tampering, the standard error of measurement is never disclosed. This may have a negative consequential validity on test takers, as students with scores at or around the cut score may or may not have passed the examination. The technical limitations inherent in state assessment results call into question the use of the results as accurate decision-making tools and challenge the reliance on standardized state or national tests as the ultimate outcome measure of education quality (American Education Research

**Item Bank**

During the process of the construction of the LOTE assessments, subject matter experts contribute items to the item bank. The process demands further scrutiny, as some of the quality issues associated with the assessment could be addressed if protocols are instituted to monitor the quality of the items before they are placed in the item bank. I recommend that each item should be coded by panels of subject matter experts according to the depth of knowledge, category of the syllabus, standards, and comments. This will make the construction of the assessment more uniform and aid in determining the reliability and validity of the assessment before it is administered.

**Categorical Concurrence**

The assessments must be aligned to the LOTE construct it purports to measure, the LOTE Checkpoint B syllabus. The policy recommendation is to expand the content coverage on the assessments to match the breadth of the content expectations in the June administration of the Spanish Comprehensive Regents Examination. Findings from this study show a variation in the categorical concurrence of the NYS Comprehensive Regents Examination. For example, categorical concurrence ranges from the lowest percentages of 51.51% in the January 2006 and June 2007 administrations to the highest percentage of categorical concurrence of 90.90% in the June 2009 administration.
Implicit in the theory of creating centralized coherent standards is that the instruction and assessment will align to those standards. For this to happen, there needs to be a sufficient degree of alignment between the curriculum, the assessment, and the standards. NYSED should also update the syllabus to include email and the Internet as modes of communication.

The NYS LOTE standards were never fully implemented. For the LOTE Modern Languages Examinations, NYSED implemented two of the four standards distributed in the mandate to all schools. NYSED needs to reexamine the standards and ensure that the State Standards are aligned to the requirements of the Common Core State Standards that NYSED adopted and is in the process of implementing.

**Depth-of-Knowledge Consistency**

The criterion-referenced test, the New York State Spanish Comprehensive Regents Examination, was reportedly constructed to assess student performance in the four skills of speaking, listening, reading, and writing. The tests under study for the years 2006-2011 bear the markers of the most recent education reform movements; namely, the standards movement and the call for increased academic rigor by raising cut scores. The tests were deployed to monitor implementation of a set of standards. In 2003, NYSED revised the LOTE Spanish Comprehensive Regents Examination to reflect the state standards, and the cut score was increased from 50% to 65%. However, the Languages Other Than English Spanish Comprehensive Regents Examinations were never studied to determine how well it aligned to the actual Spanish syllabus and standards.

Findings from this study on the alignment of the NYS Spanish Comprehensive
Regents Examination show a depth-of-knowledge consistency mostly below the cut score and an inconsistency from administration to administration of the examination, even within the same calendar year. Findings from this alignment study also show that although the cut score was increased, the depth-of knowledge-consistency was not increased. The implication is that even though students were required to earn more points to pass the examination, the cognitive rigor was easier than indicated by the cut score. The consequence of the misalignment of the level of cognitive rigor and the cut score has implications for students as well as practitioners. The implications of misalignment may mean that students are not as academically prepared to move to Checkpoint C, as implied by the passing scores; for practitioners, the misalignment creates issues of placement and the cognitive rigor in everyday instruction in the classroom. Popham (2002) posited the theory that assessment drives instruction; and therefore the low cognitive level of the assessment will have a negative washback on the instruction in the classroom, thereby lowering the cognitive rigor in classroom instruction. The negative washback then adversely impacts students’ opportunity to learn and begins to skew the possibilities for advancement in education, as the students may be neither as prepared nor as advanced as indicated by the numerical construct that the test score indicates. There must be another layer of quality control to ensure the alignment between the cut score and the categorical concurrence, the depth-of-knowledge consistency, and the range of knowledge of the items on any high-stakes standardized assessment.

**Range of Knowledge**

The narrowing of the curriculum as well as items at lower cognitive levels results
in a false assumption that test scores are going up or pass rates remaining consistent when this may not be true. The range of knowledge for the Spanish Comprehensive LOTE Examination is inconsistent from administration to administration. The results from this study suggest that blindly selecting questions, reading passages, and writing prompts that appear to match the content of the syllabus do not lead to appropriate test construction. The step beyond is to ensure alignment with the syllabus, the standards, and the cognitive demand.

**Transparency of Test Data Reports**

The NYS Report Card should be more transparent. When there is more than one administration of an assessment in a subject area, the aggregate scores should be reported for each assessment and not combined, thereby allowing educators and the general public to be aware of fluctuations of scores for individual administration and also disclosing the standard error measurements in each administration.

**Quality Control of High-Stakes Assessments**

Ultimately, the LOTE Examination was discontinued by NYSED in June 2011 and will now be written by school districts to give students the opportunity to attain an Advanced Regents Diploma. However, regardless of who funds the writing of the LOTE Examinations, the results are used to make important decisions about students. Findings from this study show that there is a variation in the cognitive level of assessments and the range of knowledge of items assessed in each administration of the Spanish LOTE Examination. With districts of varying resources and personnel who may or may not have training in psychometrics, the next wave of tests will surely vary in quality and content validity by school district across New York State.
Compliance with Federal Law

NCLB is a federal education act that challenges each state to establish a coherent assessment system based on solid academic standards. All states receiving Title I funds must present evidence of establishing a fair and consistent assessment system that is based on rigorous standards, sufficient alignment between standards and assessments, and high-quality educational results. Concerning alignment, all aspects of the state assessment system must coincide, including the academic content standards, achievements standards (linked to cut scores), performance level descriptors, and each assessment. (Taylor, Kroger, Webb, & Thacker, 2009, p. 5).

However, the federal mandate to align tests to standards and syllabi was not fully funded, nor were personnel fully trained. NYSED should resume the responsibility for the creation of the LOTE Comprehensive Regents Examination to ensure that there are sufficient and adequate quality control measures in place for an examination such as this upon which the conferral of the New York State Advance Regents Diploma is based.

Recommendations for Practice

Professional Development

Districts should provide content-specific LOTE professional development for teachers and supervisors on the syllabus, conduct alignment studies on assessments, and publish the findings of the alignment studies. Alignment should be studied across time, using several instruments to give a consistent picture of common understanding of what students learn, to provide consistent implications for instruction, to be fair to all students, and to be based on sound principles of cognitive development.
Positive Washback

“Positive washback benefits teachers, students, and administrators because it assumes that testing and curriculum designs are both based on clear course outcomes that are known to both students and teachers/testers” (Coombe, Folse, & Hubley, 2007, p. xxv). "To achieve the positive washback effect, Hasselgren (2000) maintained that the tasks should promote good learning as well as assessment activities [in which] both pupils and teachers should develop their ability to assess, based on explicit criteria" (p. 262). Well-designed performance assessments can provide strong positive washback effects, especially if they are directly linked to a particular curriculum (Hamidi, 2010).

The findings from this study demonstrate that current events such as concerns about the environment, historical events, and such appear on the assessment even though not part of the checkpoint. The recommendation is that practitioners need to incorporate more current events into the classroom setting at the correct checkpoint level while covering the breadth of the syllabus.

Recommendation for Future Research

The raw data are attached to calculate the balanced representation for the years 2006-2011 for Part 2 of the New York State Spanish Comprehensive Examination.

I also recommend conducting an alignment study from the years 2001-2005 to study changes in domain representation.

Conclusions

If ever there was a misguided policy initiative, it is that somehow high-stakes testing can drive quality into public schools by ratcheting up the consequences for not doing well on them. This approach is nothing more than the mass-legislative inspection
decried and condemned by the father of quality, W. Edward Deming (1986), who said, “Inspection to improve quality is too late, ineffective, costly” (p. 28, as cited in English & Steffy, 2001). High-stakes tests cannot monitor nor improve education. This is particularly cogent when the tests are not aligned to the policy, standards, and syllabus they purport to measure.

**Summary**

With high stakes attached to standardized examinations, it is imperative that there is a connection between the numerical score and the construct it is purporting to measure. This will facilitate better articulation, placement, and discussion among practitioners in the field, as well as create the opportunity for students to learn the intended curriculum. In addition to implementing policies that affect scoring, the assessments must be adequately funded. Personnel must be provided with training to ensure implementation fidelity of the policies and that there are adequate quality control measures that guarantee the content and construct validity of the assessments.
References


assessing: A revision of Bloom's Taxonomy of educational objectives.


Bachman, L. F. (1989). The development and use of criterion-referenced tests of


Teacher Education and Practice, 17(2), 162-181.


Educational Administration Quarterly, 30(1), 77-96.


Boykin, A. W., & Noguera, P. (2011). Creating the opportunity to learn: Moving from
research to practice to close the achievement gap. Alexandria, VA: Association for Supervision & Curriculum Development.


Contino, J. (2012). A case study of the alignment between curriculum and assessment in
the New York State earth science standards-based system. *Journal of Science Education and Technology*, 1-11.


Firestone, W. A., Schorr, R. Y., & Monfils, L. F. The ambiguity of teaching to the test:


Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and


University of California, Graduate School of Education & Information Studies.


from http://papers.nber.org/papers/w11154


Lagana-Riordan, C., Aguilar, J. P., Franklin, C., Streeter, C. L., Kim, J. S., Tripodi, S. J.,


from http://www.leg.state.mn.us/docs/2008/other/080990.pdf


173.


Moss, P. A. (1992). Shifting conceptions of validity in educational measurement:


New Mexico State Dept. of Education. (1980). *New Mexico standardized testing program report*. Santa Fe, NM: Author.

conservatism or theory-based principles? *Hispania, 66*(2), 185-191.


Ohanian, S. (1999). *One size fits few: The folly of educational standards.* Portsmouth,
NH: Heinemann.


Rowan, B., Correnti, R., & Miller, R. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. *The Teachers College Record, 104*(8), 1525-1567.


Scot, T. P., Callahan, C. M., & Urquhart, J. (2008). Paint-by-number teachers and cookie-cutter students: The unintended effects of high-stakes testing on the education of


Stewart, V. (2012). *A world-class education: Learning from international models of
excellence and innovation. Alexandria, VA: Association for Supervision &
Curriculum Development.

docview/304583251?accountid=13793 ProQuest Dissertations & Theses (PQDT)
database.


approach for evaluating a public alternative school program. Children & Schools,
33(4), 197-214.

Philosophy of science, cognitive psychology, and educational theory and practice,
(pp. 147-176). Florence, KY: Taylor & Francis Group.


selection. Alexandria, VA.: Association for Supervision and Curriculum
Development.

Sturman, L. (2003). Teaching to the test: Science or intuition? Educational Research,
45(3), 261-273.

science teaching. Greenwich, CT: Information Age Publishing.


Teacher Quality Resources LLC. (2005). Teacher quality digest. Williamsburg, VA: TRQ.

Model Schools Project.


achievement testing by ministries of education and school boards. Toronto, ON: Canadian Education Association.


Wasley, P. A., Fine, M., Gladden, M., Holland, N. E., King, S. P., Mosak, E., & Powell,


Services.


Wong, K. K., & Wang, M. C. (2002). *Efficiency, accountability, and equity issues in Title


Appendix A: Definition of Terms and Abbreviations

Alignment is the extent to which “expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do” (Webb, Horton, & O’Neal, 2002, p. 1).

Curriculum Alignment (Shepard, 1991b, 1993) is defined as a comprehensive approach to teaching and learning that goes beyond any single measure of the curriculum taught or learned. It is broadly anticipatory of any form of assessment. It is based on the doctrine of no-surprises, that is, children will not be taken by surprise regardless of the form of the assessment, as assessment is an integral part of the instructional program and not an add-on (Wraga, 1999).

Test Validity refers to the property of a test in the context of its use. There must exist a body of empirical evidence that establishes a connection between a numerical test score and the construct or trait it is supposed to measure, within the context of a particular use (Smith & Fey, 2000). That is, test validity refers to the extent to which the inferences made from a test (i.e., that the student knows the material of interest or not) is justified and accurate (Wells & Wollack, 2003). Cohen (1994) defines test validity as how well the assessment instrument has in fact measured the original objectives of the test.

Content Validity is the correspondence between curriculum objectives and the objectives being assessed. It is an aspect of construct validity that emphasizes evidence bearing on the appropriateness of the knowledge, skills, and abilities measured by an assessment. Content validity is comprised of (1) the sampling adequacy of test content, (2) the sampling adequacy of test responses, (3) the relevance of test content to a content
universe, (4) the relevance of test responses to a behavioral universe, (5) the clarity of content domain definitions, and (6) the technical quality of test items.

**Consequential Validity** is the second type of validity, referring to the way in which the assessment is used to benefit teaching and learning processes and to benefit students. It refers to the societal ramifications of testing. (Messick, 1998) Messick also defines consequential validity as an aspect of construct validity that “appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially with regard to issues of bias, fairness, and distributive justice” (Messick, 1995, p.745).

There are several other terms often associated with validity: *Face validity* (if the test looks legitimate to the respondent), and *criterion validity* (verification of the functionality of the test in comparison with another language test of equal value) are closely linked.

In New York City, as a result of new accountability benchmarks, schools with specified pass rates on Regents Examination were placed in cohorts. All schools were assigned letter grades ranging from A to F. Principals’ evaluations and bonuses became tied to test scores. When principals’ merit evaluations depend on high scores or growth on the single indicator, they in turn, put pressure on students and teachers to concentrate more of their attention on that [indicator], excluding alternative and possibly more meaningful content and teaching processes (Smith & Fey, 2000). In linguistic terms, this process is called *negative backwash* or *negative washback*.

*Washback* is an aspect of impact, or a facet of consequential validity, which has become a major area of study within the applied linguistics, especially language testing and assessment. It is sometimes referred to as *backwash* in general education (Hughes,
"Washback occurs when it is the testing instrument rather than the statement of desired learner outcomes that determines the nature of the curriculum and the course of instruction" (Valette, 1994, p. 10). The assumption underlying the research into washback is that teaching to the test where learners are provided with test-taking skills and strategies or test-driven activities or with a passive acquisition is claimed to have a negative or harmful impact on students' learning and teaching. On the other hand, when the teachers assign students activities which develop knowledge-driven, cognitive, and metacognitive skills and strategies, the washback will place a positive effect on learning as well as teaching.

A *Positive Washback* effect occurs when the assessment procedures correspond to the course goals and objectives. For instance, if a program sets a series of communicative performance objectives and tests the students using performance assessments (e.g., role plays, interviews) and personal-response assessments (e.g., self-assessments, conferences), a powerful and positive washback effect can be created in favor of the communicative performance objectives. Positive washback occurs when the tests measure the same types of materials and skills that are described in the objectives and taught in the courses (Brown & Hudson, 1998, p. 668; Hamidi, 2010).

*Reliability* is defined by Genesee and Upshur (1996) as “the freedom of nonsystematic fluctuation, or the degree to which an assessment yields consistent results.” Reliability is a test-measuring instrument which attempts to determine if a particular test is given to the same respondent on a second occasion, the results would be equal to those of the first occasion (Cohen, 1994). There are three sets of factors which
influence the reliability of a test. They are (1) test factors (ambiguity, clarity of instructions, layout of the test, length and the familiarity the respondent has with the test's format), (2) situational factors (qualities of the physical space, lighting, acoustics) and (3) individual factors (the physical health and psychological state of the respondents, their cognitive abilities, and motivation) (Cohen, 1994; (De Benedetti). An assessment is reliable when there is little difference in learners' scores or in judges' ratings across different occasions or different judges (Brindley, 2003). Reliability is based on performance, and, for an assessment to be valid, it must be reliable.

Systematic Validity (Frederiksen & Collins, 1989; Pierce, 1992; Berry, 1994; Cohen, 1994) or test impact (Bachman & Palmer, 1996), refer to factors that push measurements in the same direction. The results cannot be replicated across trials. Systematic validity is influenced by measurement-driven instruction (Popham et al., 1985; Popham, 1987) Measurement-driven instruction occurs when the testing instrument used is substituted for the curriculum.

High-stakes testing refers to testing enacted by policymakers with the intention of measuring student achievement and developing an increased effort from students, teachers, and educational administrators. Attached to these tests are consequences for failure to achieve that can go beyond the test taker (American Educational Research Association, 2000).

Standardized Testing refers to testing administered and scored in a systematic manner, with the indented outcome that the questions, conditions for administering, scoring procedures, and interpretations are reliable and are administered and scored in a prearranged, standard manner (Pearson Educational Measurement Group, 2008).
Test Preparation refers to the instruction in test taking skills and strategies provided in advance of tests for the exclusive purpose of improving scores (Kercheval & Newbil, 2002).

LOTE is an abbreviation of Languages Other Than English

NYSED is an abbreviation of the New York State Department of Education

Coding refers to deriving and developing concepts from data.

Constant Comparisons refers to the analytic process of comparing different pieces of data for similarities and differences. (Corbin & Strauss, 2008, p. 65).

Syllabus refers to a summary outline of a discourse, treatise, or course of study, or of examination requirements.
Appendix B: Request and Approval of Research

APPROVAL FOR DISSERTATION PROPOSAL

Candidate, Lisa Mars, has successfully completed all requisite requirements. This candidate’s proposal has been reviewed and the candidate may proceed to collect data according to the approved proposal for dissertation under the direction of the mentor and the candidate’s dissertation committee.

If there are substantive differences between what has been approved and the actual study, the final dissertation should indicate, on separate pages in the Appendix, the approval of the committee for those changes.

Title of Proposed Dissertation:
Content Validity of the New York State Spanish Comprehensive Regents Examination: Alignment to New York State Modern Languages Checkpoint B Syllabus and New York State Modern Languages Standards 2006-2011

Dissertation Committee:

Dr. Christopher Tienken
Mentor (Print Name)  Signature/Date

Dr. Barbara Strobert
Committee Member (Print Name)  Signature/Date

Dr. Thomas Tramglini
Committee Member (Print Name)  Signature/Date

Dr. Dennis Canterbury
Committee Member (Print Name)  Signature/Date

Approved by Seton Hall University Institutional Review Board on:

Department Chairperson  Signature/Date

Waived by IRB by: _______________________________ on this date
Appendix C: Approval of Dissertation Orals

APPROVAL FOR DISSERTATION ORALS

Candidate: ___Lisa Mars___

Dissertation Title:
Content Validity of the New York State Spanish Comprehensive Regents Examination: __Alignment to New York State Modern Languages Checkpoint B Syllabus and New York State Modern Languages Standards 2006-2011__

Academic Department: ___Department of Education Leadership Management and Policy___

Degree: ___Doctor of Education___

Committee (Please PRINT and SIGN your name):

<table>
<thead>
<tr>
<th>Mentor</th>
<th>Signature/Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Committee Member</th>
<th>Signature/Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Committee Member</th>
<th>Signature/Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Committee Member</th>
<th>Signature/Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Date Approved by Committee: Date Received by Graduate Studies Office
Appendix D: Coding Transcripts

The raw data is over 142 pages. I appended a sample below. Contact the researcher for the complete data set.

Rater 1 2007 January DOK

<table>
<thead>
<tr>
<th>Year/mth/day</th>
<th>Q#</th>
<th>DOK Level</th>
<th>DOK description</th>
<th>DOK notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007 Jan. 24</td>
<td>1</td>
<td>2</td>
<td>identify, interpret</td>
<td>Identify and summarize the main events in a narrative</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>identify patterns</td>
<td>Identify and summarize the main events in a narrative</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>Infer</td>
<td>Use context cues to identify the meaning of unfamiliar words</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>infer/use context cues</td>
<td>Identify and summarize the main events in a narrative</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>Draw conclusions</td>
<td>Support ideas with details and examples</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>2</td>
<td>interpret/identify patterns</td>
<td>Support ideas with details and examples</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>3</td>
<td>Draw conclusions</td>
<td>Identify and summarize the main events in a narrative</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2</td>
<td>infer/use context cues</td>
<td>Identify and summarize the main events in a narrative</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>2</td>
<td>Infer</td>
<td>Use context cues to identify the meaning of unfamiliar words</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2</td>
<td>interpret</td>
<td>Recall elements and details of a story, structure</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>1</td>
<td>Recite/recall</td>
<td>Recall elements and details of a story,</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>1</td>
<td>Recite/recall</td>
<td></td>
</tr>
<tr>
<td>Reading comprehension</td>
<td>13</td>
<td>1</td>
<td>Recite/recall/match/recognize</td>
<td>Recall elements and details of a story, structure</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----</td>
<td>---</td>
<td>-----------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>1</td>
<td>Match/Recognize</td>
<td>Recall elements and details of a story, structure</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>1</td>
<td>identify/ recognize</td>
<td>Support ideas with details and examples</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>3</td>
<td>Draw conclusion</td>
<td>Identify patterns, use context cues, make observations</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>2</td>
<td></td>
<td>Use context cues to identify the meaning of unfamiliar words</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>1</td>
<td>Recite/recall/match/recognize</td>
<td>Use context cues to identify the meaning of unfamiliar words</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>2</td>
<td></td>
<td>Recall elements and details of a story, structure</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>3</td>
<td>Draw conclusion</td>
<td>Support ideas with details and examples</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>1</td>
<td>Recite/recall/match/recognize</td>
<td>Recall elements and details of a story, structure</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>1</td>
<td>Recite/recall/match/recognize</td>
<td>Determine the author's purpose and describe how it affects the interpretation of a reading selection</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>3</td>
<td>Draw conclusions</td>
<td>Use context cues to identify the meaning of unfamiliar words</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>2</td>
<td>Infer</td>
<td>Support ideas with details and examples</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>3</td>
<td>Develop a logical argument</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>3</td>
<td>Develop a logical argument</td>
<td></td>
</tr>
<tr>
<td>Year/mth/day</td>
<td>Q#</td>
<td>category</td>
<td>theme</td>
<td>Topic</td>
</tr>
<tr>
<td>-------------</td>
<td>----</td>
<td>----------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>2007 Jan. 24</td>
<td>1</td>
<td>11</td>
<td>Leisure/special occasions</td>
<td>traditions and customs</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>11</td>
<td>Leisure/activities</td>
<td>hobbies/sports and other interest</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>13</td>
<td>Shopping/shoppers information</td>
<td>Advertisement</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>Personal identification/psychological characteristics</td>
<td>tastes ans interests</td>
</tr>
<tr>
<td>No.</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>13</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>Community neighborhood responsibilities / expectations</td>
<td>Shopping/shoppers information Advertisement</td>
<td>education/school life Extra curricular activities</td>
<td>Current events/Cultural aspects Arts, theater(cinema/music) Field/profession</td>
</tr>
<tr>
<td>6</td>
<td>federal information/publication. Students have to infer that because it is federal information then it is free</td>
<td>purchase mini computer</td>
<td>taking music lessons at an early age, distractors</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The table above lists various topics and their corresponding details. The entries are not complete and may require further context or clarification for full understanding.
21 15 Current events/Political, social, and economic aspects Social classes and their relations Tango was not allowed by government C

22 15 Current events/Political, social, and economic aspects Current economic issues Magazines sold at higher prices in some places A

23 15 Current events/Political, social, and economic aspects Trends in the economy Target donates money to schools with certain purchases C

24 15 Current events/Cultural aspects Special Events Carnival is an old tradition B

25 4 family Life Rapport among family members Locating lost brother C

26 15 Current events/Cultural aspects Historical and artistic sites deterioration of archaeological zones C

27 15 Current events/Cultural aspects Historical and artistic sites Murals lost color and light intensity B

28 15 Current events/Cultural aspects Historical and artistic sites Murals demonstrate a particular historical period B

29 15 Current events/Cultural aspects Historical and artistic sites Use natural productions for restoration B

30 15 Current events/Cultural aspects Historical and artistic sites Structures built during different time periods B

31 4, 7, 9, 11, B

32 1, 3, B

picture 1, 4, 11 B

9 of 15
<table>
<thead>
<tr>
<th>Year/mth/day</th>
<th>Q#</th>
<th>Standard</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007 Jan. 24</td>
<td>1</td>
<td>listening comprehension, Comprehend messages and short conversations</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Comprehend messages and conversations</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>understand the main idea and some discrete information in television,</td>
<td>radio or live presentation, Only appears from check point B</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>understand the main idea and some discrete information in television,</td>
<td>radio or live presentation, Only appears from check point B</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>understand the main idea and some discrete information in television,</td>
<td>radio or live presentation, Only appears from check point B</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>Cognates and contextual cues to derive meaning from text</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>understand the main idea and some discrete information in television,</td>
<td>radio or live presentation, Only appears from check point B</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>understand the main idea and some discrete information in television,</td>
<td>radio or live presentation, Use cognates and contextual to derive</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>meaning from text that contain unfamiliar words, expressions and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>structures</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>understand the main idea and some discrete information in television,</td>
<td>radio or live presentation,</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>understand the main idea and some discrete information in television,</td>
<td>radio or live presentation, use cognates and contextual to derive</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>meaning from texts that contain unfamiliar words, expressions and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>structures</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>Comprehend language consisting of simple vocabulary and structures</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>Comprehend language consisting of simple vocabulary and structures</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>Comprehend language consisting of simple vocabulary and structures</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>Call upon repetition, rephrasing, to derive meaning from a language</td>
<td>other than English</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>Comprehend messages and conversations when listening to peers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>Reading comprehension</td>
<td>reading and comprehend materials written for native speakers when</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>the topic and language are unfamiliar</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td></td>
<td>Read simple materials independently, but may have to guess at</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>meanings of longer more complex material</td>
</tr>
<tr>
<td>Level</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Use cognates and contextual cues to derive meaning from texts that contain unfamiliar words, expressions and structures.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Read simple materials independently, but may have to guess at meanings of longer more complex material.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>The topic and language are unfamiliar.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Comprehend the content of most texts of interest to native speakers.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Call upon repetition, rephrasing, to derive meaning from a language other than English.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Understand the main ideas and significant details of other media designed for use by native speakers. This only appears at Checkpoint C.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Comprehend the content of most texts of interest to native speakers.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Only appears at check point C.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Comprehend the content of most texts of interest to native speakers.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>The topic and language are unfamiliar.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Use cognates and contextual cues to derive meaning from texts that contain unfamiliar words, expressions and structures.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Use cognates and contextual cues to derive meaning from texts that contain unfamiliar words, expressions and structures.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Use cognates and contextual cues to derive meaning from texts that contain unfamiliar words, expressions and structures.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Write short notes, uncomplicated personal and business letters, brief journals and short reports.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Write short notes, uncomplicated personal and business letters, brief journals and short reports.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Write short notes, uncomplicated personal and business letters, brief journals and short reports.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Rater 1 2007 June DOK

<table>
<thead>
<tr>
<th>Year/mth/day</th>
<th>Q#</th>
<th>Level</th>
<th>DOK description</th>
<th>DOK notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007/June/15</td>
<td>1</td>
<td>1</td>
<td>Recite/recall</td>
<td>Recall elements and details of a story structure, such as sequence of events, character, plot and setting</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>interpret/use context cues</td>
<td>Identify and summarize events in a narrative/Use context cues to summarize major events in a narrative</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>interpret/use context cues</td>
<td>Identify and summarize events in a narrative/Use context cues to summarize major events in a narrative</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>identify patterns</td>
<td>Identify and summarize events in a narrative/Use context cues to summarize major events in a narrative</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2</td>
<td>identify patterns</td>
<td>Identify and summarize events in a narrative/Use context cues to summarize major events in a narrative</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>2</td>
<td>interpret/use context cues</td>
<td>use context cues to identify the meaning of unfamiliar words</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2</td>
<td>interpret/use context cues</td>
<td>use context cues to identify the meaning of unfamiliar words</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2</td>
<td>interpret/identify patterns</td>
<td>Use context cues to summarize the major events in a narrative</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>3</td>
<td>logical argument</td>
<td>Support ideas with details and examples</td>
</tr>
<tr>
<td>Questions in Spanish</td>
<td>10</td>
<td>1</td>
<td>Recite/recall</td>
<td>Identify and summarize events in a narrative</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>2</td>
<td>Use context cues</td>
<td>use context cues to identify the meaning of unfamiliar words(There are 3 indicators of the answer)</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>2</td>
<td>Use context cues</td>
<td>use context cues to identify the meaning of unfamiliar words</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>2</td>
<td>interpret/use context cues</td>
<td>use context cues to identify the meaning of unfamiliar words</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Learning Objective</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>----------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>Recite/recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recall elements and details of a story structure, such as sequence of events,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>character, plot and setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>Use context cues/interpret</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>use context cues to identify the meaning of unfamiliar words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>16</td>
<td>3</td>
<td>Draw conclusions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Support ideas with details and examples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>Recite/recall/match</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>use context cues to identify the meaning of unfamiliar words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>interpret/use context cues</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Answer found in subheading of reading in bold</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recall elements and details of a story structure, such as sequence of events,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>character, plot and setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>match</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recall elements and details of a story structure, such as sequence of events,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>character, plot and setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>match/recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recall elements and details of a story structure, such as sequence of events,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>character, plot and setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>realia</td>
<td>21</td>
<td>3</td>
<td>Draw conclusions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Support ideas with details and examples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>4</td>
<td>Connect/Synthesize</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analyze and synthesize information from multiple sources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>Connect/Synthesize</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analyze and synthesize information from multiple sources Students need to know</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>military time and then match the times</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>Connect/Synthesize</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Draw conclusions/ Cite evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Support ideas with details and examples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>Connect/Synthesize</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Draw conclusions/ Cite evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Support ideas with details and examples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>Use context cues/interpret</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>use context cues to identify the meaning of unfamiliar words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>2</td>
<td>Use context cues/interpret</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>use context cues to identify the meaning of unfamiliar words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q#</td>
<td>DOK Level</td>
<td>DOK description</td>
<td>DOK notes</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-----------</td>
<td>--------------------------</td>
<td>--------------------------------------</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>identify/infer</td>
<td>summarize major events</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>identify patterns</td>
<td>summarize major events</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>infer</td>
<td>context cues to identify meaning of unfamiliar words</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>infer</td>
<td>identify and summarize main event</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>draw conclusions</td>
<td>support ideas with details and examples</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>identify patterns</td>
<td>use context cues to</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>----------------------------------------------------------------------</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>draw conclusions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>support ideas with details and examples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>infer</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>identify and summarize main event</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>infer</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>identify and summarize main event</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>interpret</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>use context cues to identify the meaning of unfamiliar words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>recite/recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>recall elements and details of story</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>recite/recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>recall elements and details of story</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>match/recognize</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>recall elements and details of story</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>match/recognize</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>recall elements and details of story</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>identify/recognize</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>recall elements and details of story</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>draw conclusions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>support ideas with details and examples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>identify patterns, use context cues, make observations</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>use context cues to identify meaning of unfamiliar words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>recite/recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>recall elements and details of story</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>identify patterns</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>use context cues to identify meaning of unfamiliar words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>draw conclusions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>support ideas with details and examples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>recite/recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>recall elements and details of story</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>recite/recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>recall elements and</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
23 3 draw conclusions
determine the author's purpose and describe how it affects the interpretation of a reading selection

24 2 infer
use context cues to identify meaning of unfamiliar words

25 3 develop a logical argument
support ideas with details and examples

26 3 develop a logical argument
support ideas with details and examples

27 2 identify patterns
identify and summarize main event

28 1 recite/recall
recall elements and details of story

29 1 match/recognize
recall elements and details of story

30 1 match/recognize
describe the features of a place

31 4 create/apply concepts
conduct a project that requires specifying a problem

32 4 create/apply concepts
conduct a project that requires specifying a problem

33 4 conduct a project that requires specifying a problem
<table>
<thead>
<tr>
<th>day</th>
<th>#</th>
<th>topic</th>
<th>point</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007 Jan. 24</td>
<td>1</td>
<td>leisure/special occasions</td>
<td>tradition and customs</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>leisure/activities</td>
<td>going to the beach</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>shopping/shopping information</td>
<td>buying a car</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>personal identification/psychological characteristics</td>
<td>running for political office</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>community/neighborhood</td>
<td>federal information</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>shopping/shopping information</td>
<td>purchase mini computer</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>education/school life</td>
<td>importance of music lesson</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>current event/cultural aspects</td>
<td>field/ profession</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>current event/cultural aspects</td>
<td>art exhibition</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>leisure/available leisure time</td>
<td>what to do summer time</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>travel/transportation</td>
<td>traveling advise</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>travel/transportation</td>
<td>high speed train travel</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
<td>current event/cultural aspects</td>
<td>music award</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>personal identification/biographical information</td>
<td>return to home country</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>meal taking/food/drink,</td>
<td>eating in school cafeteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>current event/cultural aspects</td>
<td>people in the arts</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>illness and accidents</td>
<td>symptoms of illness</td>
</tr>
<tr>
<td>18</td>
<td>15</td>
<td>current event/cultural aspects</td>
<td>people in the arts</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>personal identification/psychological characteristics</td>
<td>tastes and interests</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
<td>current event/cultural aspects</td>
<td>people in the arts</td>
</tr>
<tr>
<td>21</td>
<td>15</td>
<td>current event/political, social and economic aspects</td>
<td>social classes and their relations</td>
</tr>
<tr>
<td>22</td>
<td>15</td>
<td>current event/political, social and economic aspects</td>
<td>current economic issues</td>
</tr>
<tr>
<td>23</td>
<td>15</td>
<td>current event/political, social and economic aspects</td>
<td>trends in the economy</td>
</tr>
<tr>
<td>24</td>
<td>15</td>
<td>current event/cultural aspects</td>
<td>special events</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>family life</td>
<td>rapport among family members</td>
</tr>
<tr>
<td>26</td>
<td>15</td>
<td>current event/cultural aspects</td>
<td>historical and artistic sites</td>
</tr>
<tr>
<td>27</td>
<td>15</td>
<td>current event/cultural aspects</td>
<td>historical and artistic sites</td>
</tr>
<tr>
<td>Q#</td>
<td>Standard</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>2007 Jan. 24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>comprehend messages and short conversations</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>comprehend messages and short conversations</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>understand main idea</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>understand main idea and some information in television, radio, or live presentations</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>understand main idea and some information in television, radio, or live presentations</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>understand main idea and some information in television, radio, or live presentations</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>understand main idea and some information in television, radio, or live presentations</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>understand main idea and some information in television, radio, or live presentations</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>understand main idea and some information in television, radio, or live presentations</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>use cognates and contextual cues to derive meaning from text that contains unfamiliar words, expressions, and structure</td>
<td></td>
</tr>
</tbody>
</table>
11 1 comprehend language consisting of simple vocabulary and structures
12 1 comprehend language consisting of simple vocabulary and structures
13 1 comprehend language consisting of simple vocabulary and structures
14 1 call upon repetition, rephrasing
15 1 comprehend messages and short conversations when listening to peers
16 1 read and comprehend material written for native speakers when the topic and language are unfamiliar
17 1 read simple material independently but may have to guess at meaning of longer more complex material
18 1 use cognates and contextual cues to derive meaning from text that contains unfamiliar words, expressions, and structure
19 1 read simple material independently but may have to guess at meaning of longer more complex material
20 1 read and comprehend material written for native speakers when the topic and language are unfamiliar
21 1 comprehend the content of most texts of interest to native speakers
22 1 call upon repetition, rephrasing
23 1 understand the main ideas and significant details from other media designed for use by native speakers
24 1 read and comprehend material written for native speakers when the topic and language are unfamiliar
25 1 comprehend the content of most texts of interest to native speakers
26 1 comprehend the content of most texts of interest to native speakers
27 1 read and comprehend material written for native speakers when the topic and language are unfamiliar
28 1 use cognates and contextual cues to derive meaning from text that contains unfamiliar words, expressions, and structure
29 1 use cognates and contextual cues to derive meaning
30 1 use cognates and contextual cues to derive meaning
31 1 Write short notes, uncomplicated personal and business letters, brief journals and short reports
32 1 Write short notes, uncomplicated personal and business letters, brief journals and short reports
reports

33 1 Write short notes, uncomplicated personal and business letters, brief journals and short reports
Appendix E. Webb Alignment Tool


http://wat.wceruw.org/index.aspx
Appendix F: New York State Checkpoint Syllabus for Modern Languages


Appendix G: New York State Modern Language Standards

Appendix H: New York State Spanish Comprehensive Regents Examinations 2006-2011

http://www.nysedregents.org/Spanish/

Appendix I: Workshop Materials for Panelists

For the alignment review, I would like you to perform several tasks to evaluate test items as compared to the New York State mandated Checkpoint B syllabus, Standards, and Webb’s Depth of Knowledge. The tasks will be performed individually with read back and discussion after each question. You will enter your ratings and comments on the excel sheet provided on your laptop. Please ask me if you have any questions.

There are two folders in front of you. The first folder contains New York State Spanish Comprehensive Regents Examinations for the years 2006-2011. Both January and June administrations are included as well as teachers’ dictation and answer key.

In the second folder, you will find a Webb’s Depth of Knowledge chart, New York State Checkpoint B syllabus, and the New York State Modern Language Standards. A copy of the topics of the Checkpoint B syllabus is also included for quick reference.

Begin by entering the year and date of the examination on which you are currently working.

Using the Depth of Knowledge chart, with descriptions, rate each item to the appropriate depth-of-knowledge category, based on the cognitive process expected of students to devise the answer. Enter the DOK level and comments in the excel spreadsheet. Remember that cognitive complexity is related to difficulty but that these terms are not synonymous. If you find that an item requires several different tasks of varying complexity, then indicate the highest DOK level, Level 4 (Extended Thinking).
Higher-level thinking is central to Level 4. The criterion for this level is a multi-paragraph composition that demonstrates synthesis and analysis of ideas or checkpoint themes. There is evidence of awareness of purpose and audience. Students are expected to create compositions in the target language that demonstrate a perspective on the themes of the checkpoint based on the prompt.

After each rating, there will be a group discussion to achieve majority agreement on the ratings.

Repeat the steps for standards and syllabus ratings. Include a comment or descriptor for each rating.

At the end of each rating session, email your file so that there is a back up of the data.

Example of a blank rating sheet

<table>
<thead>
<tr>
<th>Year/mth/d</th>
<th>Q#</th>
<th>DOK Level</th>
<th>DOK description</th>
<th>DOK notes</th>
<th>category #</th>
<th>category/theme</th>
<th>Topic Notes</th>
<th>Checkpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year/ mth/d ay</td>
<td>Q#</td>
<td>DOK Level</td>
<td>DOK description</td>
<td>DOK notes</td>
<td>category #</td>
<td>category /theme</td>
<td>Topic</td>
<td>Item Notes</td>
</tr>
<tr>
<td>---------------</td>
<td>----</td>
<td>-----------</td>
<td>-----------------</td>
<td>-----------</td>
<td>------------</td>
<td>----------------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Content Validity of the New York State Spanish Comprehensive Regents Examination:
Alignment to New York State Modern Languages Checkpoint B Syllabus and New York State Modern Languages Standards 2006-2011

Lisa Mars
Seton Hall University

Recommended Citation:
Mars, Lisa, "Content Validity of the New York State Spanish Comprehensive Regents Examination" (2013). Dissertations. Paper_______.
http://scholarship.shu.edu/dissertations/_________
APPROVAL FOR SUCCESSFUL DEFENSE

Doctoral Candidate, Lisa Mars, has successfully defended and made the required modifications to the text of the doctoral dissertation for the Ed.D. during this Spring Semester 2013.

DISSERTATION COMMITTEE

Mentor:
Dr. Christopher Tienken

Committee Member:
Dr. Barbra Strobert

Committee Member:
Dr. Thomas Tramglini

Committee Member:
Dr. Dennis Canterbury

The mentor and any other committee members who wish to review revisions will sign and date this document only when revisions have been completed. Please return this form to the Office of Graduate Studies, where it will be placed in the candidate's file and submit a copy with your final dissertation to be bound as page number two.