EXAMINING ALIGNMENT INDICES' VALIDITY AS MEASURES OF TEST CONTENT REPRESENTATIVENESS

By

Anne Traynor

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods—Doctor of Philosophy

2014

ABSTRACT

EXAMINING ALIGNMENT INDICES' VALIDITY AS MEASURES OF TEST CONTENT REPRESENTATIVENESS

By

Anne Traynor

Alignment index values are often presented as evidence that test content is representative of the performance domain defined by a written curriculum. Alignment measures bear on the validity of state achievement test score interpretations, and on test fairness. While alignment reports have been used to document test content distribution, and to generate recommendations for test form improvement that appear sensible to assessment professionals (Schafer, Wang, & Wang, 2009), there is little external evidence that alignment index values are valid quantitative measures of tests' content representativeness.

Using eleven states' mathematics achievement test item and Surveys of Enacted Curriculum (SEC; Porter, 2002) content analysis data, I examine external validation evidence for the curriculum emphasis measures underlying the coarse-grained SEC test-curriculum alignment index. I then use fractional logit regression models (Papke & Wooldridge, 1996) to assess the relationship between state-level test-curriculum alignment and proportion-correct item difficulty on corresponding test items, controlling for other state and test item characteristics that may affect achievement. This study focuses on evaluating the validity of alignment indices as measures of test-curriculum correspondence, rather than on establishing cutoff criteria on the measures or comparing alignment data collection protocols, although these are also important issues.

I find that the content analysis proportions that summarize SEC alignment panelists' judgments about curriculum objectives' topics and cognitive demand requirements seem to relate in expected ways to other measures of state curricular emphasis in Grade 4, providing weak external validation support for the mathematics curriculum content analysis data. However, there is no evidence of a statistically or substantively significant relationship between an alignment measure based on SEC coarse-grained content analysis data, and test item difficulty in Grade 8, regardless of the extent to which a particular item's content type is emphasized by the curriculum. I conclude that although external validation studies for the SEC alignment index have tended to focus on the relationship between curricular alignment and test performance, other types of evidence may yield clearer conclusions, and perhaps be more crucial for demonstrating these indices' validity. Specifically, I suggest future research to evaluate the content classification schemes implemented by popular alignment methods, and to establish the reproducibility of overall and item-level alignment results across independent panels of qualified experts trained by different facilitators.

ACKNOWLEDGEMENTS

I was fortunate to have the support of my dissertation advisor, Dr. Mark Reckase, and my committee members, Drs. Robert Floden, Richard Houang, and Raven McCrory. Their willingness to entertain my proposal, and to ask questions and offer suggestions contributed markedly to the finished product. I appreciate the assistance of Dr. Barbara Schneider, and Michelle Chester in the Office of the Hannah Chair, who maintained and provided access to the NAEP data used in this dissertation.

I also wish to acknowledge others who contributed to my development as a scholar during my studies. I am particularly appreciative of the support and involvement of my academic advisor, Dr. Tenko Raykov, in providing feedback and advice as I progressed through my degree program. Dr. Alexander von Eye offered constructive criticism on drafts of some of my earliest work—I am grateful for his encouragement. Dr. Spyros Konstantopoulos provided thoughtful advice as I was completing this dissertation and searching for employment. My fellow students, especially Cheng-Hsien Li and Hyesuk Jang, contributed to my learning throughout my studies, and I am glad to have had their company. Finally, I am thankful for the good counsel of my family and friends, who have supported me as I have completed this dissertation, and in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AME	Average marginal effect
BIC	Bayesian information criterion
CI	Confidence interval
ESEA	Elementary and Secondary Education Act of 1965
IEA	International Association for the Evaluation of Educational Achievement
NAEP	National Assessment of Educational Progress
NCTM	National Council of Teachers of Mathematics
OTL	Opportunity to learn
TIMSS	Trends in International Mathematics and Science Study (formerly Third International Mathematics and Science Study)
SEC	Surveys of Enacted Curriculum
USED	US Department of Education
VIF	Variance inflation factor

CHAPTER 1: INTRODUCTION OF THE RESEARCH QUESTIONS

In the context of state achievement testing for Grades K–12, alignment can be defined as the degree of content correspondence between a test instrument used to measure students' achievement in a specific subject area, and a state's curriculum documents for that subject at a given grade level (Webb, 1997). There are two sources of imperfect alignment: some content in the curriculum is not tested, or a test includes some material that is not in the curriculum (La Marca, 2001). Like setting passing scores on tests, or scoring performance tasks, all alignment review methods require human judgment (Rothman, 2003; Crocker & Algina, 1986). Alignment may be reported using qualitative descriptions, or quantitative indices, which are the focus of this paper. A variety of indices have been proposed to quantify the content correspondence of particular test-curriculum document pairs, many of which are based on the topic-by-cognitive complexity classification tables often used to guide item writing and test form assembly. Alignment measures indicate the fidelity of a test to representative sampling from a particular curricular domain (McMaken & Porter, 2012; see also Guion, 1977, regarding content validation evidence).

Throughout this paper, unless otherwise stated, I apply the narrowest definition of alignment found in the literature, as a characteristic of particular test-curriculum document pairs, although the term alignment has sometimes been used more broadly, to refer to correspondence between an entire assessment system—all data collection instruments for all grade levels, and their administration, scoring and score reporting—and a given curriculum (e.g., Webb, 1997). Judgments about the overall quality of an assessment system, including accessibility (La Marca, Redfield, & Winter, 2000) and coherence (Rabinowitz, Roeber, Schroeder, & Scheinker, 2006),

are thus beyond the scope of my alignment definition, as are judgments about the quality or accuracy of item content (Plake, Impara, & Buckendahl, 2004).

Establishing evidence of test relevance to, and representativeness of, the intended achievement domain is a crucial first step in validating test score interpretations (La Marca, 2001). Alignment indices are among many types of test content evidence that are relevant for validation (Martone & Sireci, 2009). They have been used as validation evidence for tests that gauge examinees' accomplishment of formal, written curriculum expectations, primarily state-administered educational achievement tests, although in principle they could also be used as evidence for licensure or employment exams that sample content domains based on job analysis. Alignment measures bear on the validity of interpretation of students' test scores as measures of curriculum goal attainment (La Marca, 2001) because the true meaning of a test's score scale is derived from the observed content domain (Martineau, Paek, Keene, & Hirsch, 2007) represented by the test items. If a test is intended to measure curricular achievement, the observed content domain will differ from that intended domain, and the true meaning of scores will differ from their intended interpretation, to the extent that the test is misaligned with the curriculum (Martineau et al., 2007). Because a test's degree of alignment to a specific curriculum document can alter test score interpretations, following the validation framework described by Kane (2006, 2013), alignment also secondarily affects the soundness of any proposed test score uses that require those particular interpretations.

As well as being salient to educational measurement, alignment between state achievement tests and curriculum documents is important for educational practice because teachers rely on both as indicators of the intended curriculum—state policies on instruction (Porter, 2002). Alignment can be viewed more broadly as the extent to which "expectations and

assessments are in agreement and serve in conjunction with one another to guide the [educational] system toward students learning what they are expected to know and do" (Webb, 1997, p. 3). If an external test does not correspond to the curriculum, and incentives are attached to the scores, teachers will tend to align their instruction to the content and features of the assessment, rather than to the intended, written curriculum (Koretz, 2008). Developing methods to quantify the degree of match between a formal curriculum document and a test is necessary to foster adequate correspondence so that educators receive a consistent message about the intended curriculum. Although I recognize the importance of alignment between test content and instruction, also, as a fairness issue for examinees (e.g., Resnick, Rothman, Slattery, & Vranek, 2004), and so an ethical and legal issue for test score users (La Marca, 2001; Phillips & Camara, 2006), since this paper focuses on alignment between test content and a curriculum as an accuracy issue for score interpreters, it does not directly address ethical or legal concerns. It will, however, explicate and account for the relationship between test-curriculum and test-instruction alignment when it is pertinent to my argument.

This study has two main purposes. First, I use empirical test-curriculum alignment data from 11 states, together with test item difficulty data and teacher content emphasis data for their student populations, to evaluate the external validity of the curriculum content analysis proportions underlying a popular alignment index as measures of curricular emphasis. Second, I use components of the index to test hypotheses about the relationship between test-curriculum alignment, and state mean test item performance. If the implemented curriculum—instruction—follows the written curriculum, test-curriculum alignment, in conjunction with content emphasis (Gamoran, Porter, Smithson, & White, 1997), should be positively associated with student achievement on relevant tests (Crocker, Miller, & Franks, 1989). Evidence of such an

association after controlling for other factors that affect students' performance on achievement items could support the validity of existing alignment indices as measures of test-curriculum correspondence.

1.1 Test-Curriculum Alignment Evidence as a Federal School Accountability Testing System Requirement

Educational accountability systems consist of the use of test scores or other measures by government agencies to monitor the educational status and progress of, and to determine the distribution of rewards and sanctions to, schools or individual students (Linn, Baker, & Betebenner, 2002). The No Child Left Behind Act of 2001 amended and reauthorized the federal Elementary and Secondary Education Act of 1965 (ESEA). The 2001 emendation of this public law will henceforth be referred to as the "amended ESEA," or simply as the "ESEA" for brevity. The amended ESEA compelled states to develop detailed curriculum documents outlining learning expectations for reading and mathematics in each grade 3–8, and assessments aligned to those curricula, which were to be administered in public schools receiving federal ESEA education funding. State departments of education were required to use differences in school mean assessment results across years to identify public schools that produced inadequate gains in test scores, to compel these schools to adopt reform strategies, and to evaluate the effectiveness of the selected reforms (Linn et al., 2002). The amended ESEA effectively created a system of rewards and sanctions for public schools, pressuring teachers and administrators to produce high test scores. The full requirements of the amended ESEA were first effective for the 2005–2006 school year.

Any alignment method used to evaluate state achievement tests under the ESEA must, at a minimum, evaluate the degree of both content and cognitive complexity correspondence

between tests and curriculum documents (Davis-Becker & Buckendahl, 2013; La Marca et al., 2000; US Department of Education [USED], 2004). The three most commonly-used alignment methods (Martone & Sireci, 2009) are those developed by Webb (1997), by Achieve, Inc. (Resnick et al., 2004), and by Porter and colleagues based on the Surveys of Enacted Curriculum (SEC; Porter, 2002). Davis-Becker and Buckendahl (2013) contend that, because some language used in USED's (2004) initial guidance to states regarding ESEA alignment requirements was terminology specific to the Webb method, some state departments of education interpreted this to suggest that the Webb method was the "correct" or preferred alignment method. They argue that, given the limited research available on alignment methodology, neither policymakers nor measurement practitioners have enough information to allow definitive comparison of methods, or identification of preferred methods.

To judge the adequacy of assessment systems proposed by each state for ESEA compliance, USED relied on independent panels of assessment professionals to review each state's submission and complete an advisory report, summarizing validity evidence collected by each state, and noting any additional evidence that should be required before approval of a particular assessment system. Schafer, Wang, and Wang (2009) found that as a consequence of the review panels' evaluations, 23 states were asked to submit results of test-curriculum alignment studies—the most common type of required omitted evidence cited in states' assessment system proposal decision letters from USED. States' peer review reports implied that alignment results were interpreted by the review panels as evidence that item responses would be influenced by the cognitive processes intended at each grade level, and the decision letters indicated that USED required these results as validation evidence to obtain assessment system approval (Schafer et al.).

1.2 Integration of Alignment Review into the Test Development Process

Alignment between a test and a particular curriculum document is different than correspondence of a test to its specifications. In principle, a test could be judged to adequately match its specifications, but to be entirely unrelated to a curriculum that it was intended to measure. In practice, since test specifications for state achievement tests are developed with reference to a particular curriculum document, the alignment between a test and its specifications bears on the alignment between the test and the target curricular domain. Both test blueprints and test instruments should be aligned to the content standards (Martineau et al., 2007). However, typically judgments about whether a test is sufficiently aligned to its specifications are left to the test developer (Buckendahl, Plake, Impara, & Irwin, 2000), while reviews of alignment between the test and its relevant curriculum are conducted by independent panels of subject-matter experts.

Generally, sequential development of a written curriculum, test specifications and test items, in that order, is more likely than concurrent or differently-ordered development to produce aligned test-curriculum combinations (Webb, 1997). States typically conduct test-curriculum alignment studies only following major curriculum revision, test modification, or changes in test passing scores (La Marca, 2001; Wyse & Viger, 2011). However, ideally, alignment review should occur regularly during the instrument development and revision cycle (La Marca, 2001). Alignment reviews often detect sections of the curriculum that are over- or underrepresented on assembled test forms or in the item pool (Martineau et al., 2007). Although further item development is unlikely to fully address these gaps, repeated formal alignment review of the items for an annual testing cycle is uncommon and would be costly (La Marca, 2001; Martineau

et al., 2007). To improve the efficiency and coherence of the test development process for state assessment systems, Martineau et al. (2007) suggested that alignment should be monitored during the early phases of test development, particularly during item writing, and that formal alignment review of assembled test forms or the item pool should be combined with the phase of item quality review by subject-matter experts. Interpretation of other types of validity evidence, including reliability estimates and factor analysis results, should then be informed by the alignment results (La Marca, 2001)

In practice, alignment review is typically conducted for specific test forms after they are assembled (Wyse & Viger, 2011), but often only a single form is analyzed for each grade level even when multiple forms exist (Polikoff, 2012a, p. 361). The practice of judging and reporting alignment for only one of many test forms is problematic because some nominally-parallel forms may be better aligned to a curriculum than others, as demonstrated among New York Regents Exam forms (Liu & Fulmer, 2008). Schafer et al. (2009; see also Porter, 2002, p. 13) contended that alignment results for a single test form should not be judged as sufficient for ESEA compliance by states that utilize multiple test forms, although current interpretations of the law seem to have deemed alignment results from one form to meet the minimum evidence requirement. La Marca (2001) pointed out that the extent to which individual test instrument, and assessment system, alignment to a curriculum are important depends on how information from the system will inform decision-making. If decisions about students, teachers or schools will be made based on single test scores, individual test-curriculum alignment evidence is most critical to the validity of score interpretation and the soundness of the decisions. If decisions will be made based on multiple measures of curricular attainment (e.g., routine classroom assessment

scores; large-scale summative test scores), evidence of overall assessment system alignment to the curriculum may be needed.

1.3 The Surveys of Enacted Curriculum (SEC) and Other Processes for Judging Alignment

Arguing for more standardized evaluation of the quality of content domain sampling in test instruments, Guion (1977, p. 7) held that "the notion of content relevance is a quantitative one, even if we currently lack the means of measuring it." Others view alignment as fundamentally a qualitative issue: "evaluating the quality of alignment requires a holistic judgment. The purposes of the assessments and standards, their use in guiding instruction and decision-making, and other contextual information must be considered in judging whether the degree of alignment is sufficient" (La Marca et al., 2000, p. 24; see also Beck, 2007). Modern alignment procedures integrate these perspectives, generating both numeric indicators and narrative depictions, which are combined into an overall evaluation (e.g., Flowers, Wakeman, Browder, & Karvonen, 2009), although the extent to which the summative evaluation relies on the numeric indicators, and treats them as quantitative, depends on the method.

Because modern alignment methods were developed largely without reference to previous conceptualizations of test-domain content correspondence in the validation literature, and each reflects different beliefs about what test or item properties constitute "good" alignment—the methods define alignment differently—isolating important dimensions along which they can be compared is difficult. Bhola, Impara, and Buckendahl (2003) classified alignment methods by their complexity. "Low complexity" alignment methods posit a content model based on a simple ordinal rating of content match between document items (p. 22). This simple model underlies all other alignment methods. "Moderate complexity" methods characterize and rate two distinct aspects of each document task's content: the topic, and the

relative cognitive demand of succeeding on that task (p. 22). In accord with methods used to develop test specifications and write items (e.g., Haladyna, Downing, & Rodriguez, 2002), most alignment procedures adhere to, minimally, such a two-dimensional content model (La Marca, 2001). "High complexity" alignment methods consider further dimensions possibly relevant to judging the coherence of a curriculum-assessment system, such as the correspondence between test administration and instructional conditions, or between the types of performances elicited by test items and those actually stated by the curriculum objectives (Bhola et al., 2003). Among alignment methods that have been used for ESEA compliance (Martone & Sireci, 2009), which will be described in the sections that follow, Bhola et al. label the SEC method as moderate complexity, and the Webb and Achieve methods as high complexity.

1.3.1 The SEC Alignment Method

SEC alignment reviews implement a matching-type alignment method (D'Agostino et al., 2008) in which judges separately match the items composing a particular test, and the curriculum objectives the test is intended to measure, to a two-dimensional content classification matrix (Porter, 2002). Unlike indices of item-objective congruence and some modern alignment methods that pair items directly with objectives, the SEC method matches items and objectives only indirectly, through their assigned positions in the content matrix. The topic-by-cognitive demand matrices underlying the SEC have their "conceptual origin" in the late 1970s to mid-1980s work of the Content Determinants Group of the Institute for Research on Teaching at Michigan State University (Porter, 2002, p. 12). An exemplar of Content Determinants Group's work is a study by Freeman et al. (1983), which used a three-dimensional 1,260-cell topic-by-cognitive-demand-by-mathematical-operation taxonomy to classify the content of mathematics textbooks. The topic and cognitive demand classification schemes used during SEC alignment

reviews have been further informed by more recent analyses of K–12 textbooks, standardized tests, state and district curriculum documents, and curriculum recommendations of national professional educators' organizations (McMaken & Porter, 2012). Revisions of the initial content taxonomy occurred in 2004, and between 2006 and 2007 (Polikoff, 2012a, p. 347).

Experts in the subject matter of a given educational document (test or curriculum) are recruited as SEC alignment panelists (Polikoff, 2012a). They may or may not have previous knowledge of any specific curricula to be analyzed. (If a state has maintained a consistent curriculum over time, and SEC content analysis results for the curriculum are already available, it would not need to be re-analyzed [Porter, 2006]). Panelists undertake training by a moderator following a standard protocol before beginning operational coding. While the training process "is largely consistent across groups," as would be expected some "variation in the level of content expertise and experience" across panels does exist (Smithson & Collares, 2007, p. 3). After training, judges code document content independently of one another (Polikoff, 2012a), although they are given the opportunity to discuss any "flagged" items, usually a small fraction of the total, that "cause confusion for the coding process" (Porter, Polikoff, Zeidner, & Smithson, 2008, p. 4). Panelists are permitted to change their initial coding, but are not required to reach consensus or otherwise encouraged to reconcile their judgments.

Polikoff (2012a) details the coding procedures that are used by SEC judges. Test items or objectives—the most specific level of curriculum goal—are coded by both their topic and cognitive demand, which locates their positions in the content matrix. Judges are directed to match each objective to between 1 and 6 cells, and each item to between 1 and 3 cells, in the content matrix (p. 347). Although both types of task statements can cover multiple topics and cognitive demand levels, because test items are typically narrower in scope than objectives, their

classification process is more tightly restricted (Polikoff, 2012a, p. 347). In coding objectives, judges must accurately interpret the meaning of each of these performance statements that was intended by policymakers. Likewise, in coding items, judges must determine the content needed to correctly answer each, and infer "the most likely approach" that students will use in responding, which is a significant challenge (Porter, 2006, p. 147).

Once panelists have classified each task statement to cell(s) of the content matrix, to generate a standardized data matrix for each panelist for each document analyzed, the maximum score value for each test item is divided equally among content cells to which the item was matched. For example, if a two-point constructed response item is matched to three cells, each cell would be assigned two-thirds of a point (Polikoff, 2012a, p. 347). Each curriculum objective is assumed to have unit weight, which, similarly, would be divided across cells to which that objective was assigned. The weighted item or objective counts in each classification matrix are then converted to proportions by dividing each cell by the maximum test score, or total number of objectives in the curriculum document, respectively. These transformations yield a matrix of content proportions for each judge. The proportion in each cell is an estimate of the relative emphasis of that content category on the test, or in the curriculum (Porter, 2002). All panelists' matrices of content proportions for a given educational document are, finally, averaged across panelists to produce the aggregate content analysis results for that document (Polikoff, 2012a, p. 347). (Alternatively, if measuring test-instruction or instruction-curriculum alignment was the goal, in this step, matrices of content emphasis proportions for particular instructors could be derived from SEC teacher survey data [e.g., Porter, 2002].)

Aggregate summary data from the content coding process takes the same form for both curriculum and test documents, as proportions of total content in each cell of the matrix

(Polikoff, 2012a, p. 347). For any test and curriculum that have been rated, an overall test-curriculum alignment index can be computed. Given two content-by-cognitive demand matrices, reasonable measures of alignment would capture the degree of equality, or association, between proportions in the corresponding cells of the two matrices (Porter, 2002). Porter (2002) suggested two potential alignment indices for SEC data, only one of which has been used in published research. The SEC alignment index is

$$1 - \frac{\sum_i \sum_j \left| \pi_{X_{i,j}} - \pi_{Y_{i,j}} \right|}{2}$$

where $\pi_{X_{i,j}}$ denotes a cell proportion in matrix *X* and $\pi_{Y_{i,j}}$ denotes the corresponding cell proportion in matrix *Y*. This index is bounded between 0 and 1, inclusive, with higher values indicating better alignment between the coded assessment and curriculum document. The index is the sum of cellwise intersections between the two content matrices (Porter, 2006). It operationally defines a test as "aligned" with a particular curriculum to the extent that the proportion of test items in each content matrix cell is equivalent to the proportion of objectives in that cell (Porter, 2002).

An alternative alignment index proposed by Porter (2002, p. 6), since both the rows and columns of the matrix are treated as nominal rather than ordinal, is to simply compute the correlation between corresponding cells of matrices *X* and *Y*. A third possibility, analogous to a method implemented by Polikoff and Porter (2012) for instruction-curriculum alignment, in case the reliability of judges' item or objective cognitive demand classifications is in doubt, is to compute alignment based on topic coverage proportions only, collapsing over cognitive demand categories. It is also possible to generate other statistics using SEC content analysis data; for example, Polikoff (2012b, p. 285) describes computation of an index for "focus," the extent to

which the content of a state curriculum document is concentrated in certain topic-cognitive demand cells, rather than diffuse. Distributions of topic and cognitive demand classifications for items or objectives, by rater, can be produced, and diagnostic information regarding particular sources of misalignment can be acquired from content emphasis graphics illustrating proportions of test or curriculum material in various content cells (Porter, 2002).

1.3.2 Other Methods

Webb (1997, 2007) developed the first modern alignment method. In the Webb method, following training, panelists rate curriculum objectives on a four-point ordinal scale of cognitive demand, using verbs in the objectives to distinguish cognitive demand levels. After panelists render independent judgments, a facilitator, who also participates in rating, leads them in reaching consensus about each objective's demand level (Webb, Alt, Ely, Cormier, & Vesperman, 2005). Reviewers then independently match items to objectives based on their content topic, and rate items on cognitive demand, combining rating- and matching-type alignment methods (D'Agostino et al., 2008). Each item can be matched with up to 3 objectives. The Webb (2007, p. 7) method operationally defines alignment as composed of four aspects, each of which is measured by an index: "depth of knowledge," "balance of representation," "categorical concurrence," and "range of knowledge." (Webb's original framework [1997] outlined a very high complexity [Bhola et al., 2003] method that appraised additional test and curriculum properties related to assessment system coherence [Rabinowitz et al., 2006]; however, these other characteristics have rarely been measured during applications of the Webb method [Martone, 2007], and are beyond the scope of Webb's more recent alignment recommendations.)

Webb's (1997) depth of knowledge index indicates, for each curriculum strand, the proportion of matched items that require cognitive demand at or above the level of their corresponding objectives, averaged across panelists. It measures how well the test matches the intended, or a more intellectually-challenging, curriculum. The balance of representation index for each curriculum strand ranges between 0 and 1, and is based on the difference between the proportion of assessed objectives in the strand represented by a particular objective, and the proportion of items assigned to the strand that are matched to that objective. Its computation assumes that curriculum goals have more than one level of detail, and that the most specific statements, objectives, are comprehensive and equally weighted (Bhola et al., 2003, p. 24). As objectives are measured by equal numbers of items, the value of the balance of representation index will be near 1. If the index is near 0, "then either few objectives are being measured, or the distribution of items across objectives is concentrated on only one or two objectives" (Bhola et al., 2003, p. 24). Two further indices, categorical concurrence and range of knowledge, give average counts of items matched each content strand or specific objective, respectively, each of which is compared with a minimum criterion value. Because all index computations include data from every item-objective match, regardless of whether a panelist deemed the item and objective to require equal cognitive demand, or whether the item distribution for a particular content strand was balanced, interpretations of the four indices are partially confounded (Webb, Alt, Ely, Cormier, & Vesperman, 2005). While the Webb alignment method yields separate index values for the four alignment aspects for each content strand in a curriculum document, which are usually treated as distinct measures in final alignment evaluation reports, a technique to combine the indices into a single measure has been proposed (Brown & Conley, 2007).

Resnick et al. (2004) describe Achieve, Inc.'s alignment evaluation method. Achieve's alignment evaluations stem from a broader conception of alignment than either the SEC or Webb (2007) methods: "how well all policy elements in a system work together to guide instruction and, ultimately, student learning" (Resnick et al., 2004, p. 4). To minimize the rater training requirement, Achieve alignment reviews are typically conducted by an experienced external panel, rather than by assessment stakeholders. Initial confirmation or revision of the matches between items and objectives indicated by a test developer's test specifications is typically conducted by single senior reviewer. These pre-confirmed item-objective matches are assumed as a basis for the panel's alignment review. Once the test specifications table is confirmed, panelists examine each item and its relation to the objective designated in the (revised) test specifications. The product of an Achieve alignment review is an evaluation report drawing on information collected from several rating scales, one index, and qualitative assessments of test features.

Achieve's "content centrality" rating scale ranks the extent to which each item's content matches the content of its corresponding objective (Resnick et al., 2004, p. 6). The "performance centrality" rating scale ranks the extent to which the response process required by each item is consistent with the verb in the corresponding objective (p. 6). The "range" index is the proportion of curriculum objectives with content that is reflected by at least one test item (Resnick et al., 2004, p. 7), a traditional alignment index (Crocker et al., 1989). The "source of challenge" factor determines whether item performance is likely to be unduly influenced by item characteristics that are irrelevant to attaining the corresponding behavioral objective (p. 7). The "level of challenge" factor describes the anticipated difficulty of the set of items measuring a particular curriculum strand for the examinee population; its evaluation assumes that items

matched to each goal should be distributed across challenge categories in a grade-level-appropriate manner (p. 7). The "balance" aspect of alignment is a qualitative evaluation of the extent to which objectives subsumed by a particular broad goal or content strand are well-represented by their corresponding test item set, with appropriate emphasis on content that reviewers judge to be important at that grade level (p. 7). Evaluating the balance and level of challenge alignment aspects requires reviewers to make explicit value judgments about the nature of content that should comprise states' K–12 curricula and achievement tests (Rothman, 2003); both, overall, should be "sufficiently challenging" for students (Resnick et al., 2004, p. 6) and should emphasize the "more important" content at each grade level (p. 8).

1.3.3 Rationale for Studying the SEC Test-Curriculum Alignment Index

While alignment reports "have been used successfully to document . . . content representation, as well as to generate recommendations for improvement that seem to make sense" to other assessment professionals, and to those associated with a given testing program (Schafer et al., 2009, p. 182), there is little external evidence that alignment index values are valid quantitative measures of tests' content representativeness. If assumptions underlying computation of the indices (or, choice of their cutoff criteria; Webb, 2007) are not reasonable in practical situations, it is unlikely that the indices can support accurate conclusions about the degree of test-curriculum alignment. Davis-Becker and Buckendahl (2013) recommended seeking external validity evidence to evaluate an alignment panel's conclusions based on "connections to results of similar studies or other types of information" (p. 30; see also Crocker, Miller, & Franks [1989] for a similar recommendation regarding traditional content validation measures). However, because the alignment criteria applied by various methods differ in number, content, and interpretation, representing considerably different definitions of

"alignment," different methods cannot be expected to yield consistent decisions about alignment for a particular test-curriculum combination. Thus, it is unclear what conclusions could be drawn from an empirical study of multiple alignment methods that could not be inferred from existing thorough comparisons of the methods' criteria and procedures (e.g., Martone & Sireci, 2009; Vockley, 2009). Instead, I focus on evaluating the assumptions of a single type of alignment index, and then discuss the extent to which the findings are generalizable to other alignment indices.

The SEC test-curriculum alignment method meets the criteria to be used for ESEA school accountability testing alignment reviews (USED, 2012), and is sometimes employed in education research (e.g., Kurz, Elliott, Wehby, & Smithson, 2010) although perhaps less frequently used by state testing programs than the Webb method (Davis-Becker & Buckendahl, 2013). It does not require use of proprietary training materials, and the data produced is open to public scrutiny. The national SEC data repository allows partnering states and local districts to compare their curricula to those of other states or professional organizations, or to the assessment frameworks of nationwide tests (Porter, McMaken, & Blank, 2011).

Previous alignment study results generally suggest achieving adequate overall representation of curriculum objectives by a test item set is a more serious problem for state achievement test developers than devising unbiased items that measure at least one objective in the curriculum (Resnick et al., 2004; Webb, 1999). The SEC method produces a single measure of overall curriculum domain representation by a particular test, which simultaneously considers each item or objective's topic and cognitive demand. The SEC alignment index operationalizes a precise alignment definition that is similar to definitions presented by measurement theorists (Guion, 1977, p. 7). It does not reflect judgments about the educational or societal value of

particular curriculum objectives or test items, unlike the Achieve alignment components and Webb depth of knowledge index (Rothman, 2003, p. 23; Webb, 2007), and does not penalize tests that sample from the curriculum domain, rather than exhaustively assessing each objective, unlike the Webb alignment criteria (Bhola et al., 2003). Furthermore, the separate coding of test items and curriculum objectives to a common matrix may improve comparability of results across reviews since curriculum documents' specificity does not dictate the level of detail at which topic and cognitive demand categories are defined (Porter, 2006, p. 149), and may reduce the bias that would be likely to result if items and objectives were directly matched (Anderson, 2002). Since theory provides some support for interpretation of the SEC alignment index, this paper seeks external empirical evidence to contribute to validation.

1.4 Purpose of Study and Research Questions

Kane (2013, p. 13) emphasized that test score interpretation validation arguments should focus on identifying, detailing, and evaluating inferences and assumptions that are "most questionable *a priori*." Alignment index validation arguments likewise should concentrate on testing questionable assumptions. Measurement theorists have characterized alignment as the extent that tests' content emphasis matches the emphasis of a relevant curriculum (La Marca et al., 2000; Poggio, Glasnapp, Miller, Tollefson, & Burry, 1986). Quantifying curriculum content emphasis is recognized to be a complicated issue (Crocker et al., 1989), perhaps without a universal best solution. The calculations for both the SEC and Webb balance of representation test-curriculum alignment indices assume that curriculum objectives are intended to receive equal coverage, so that unweighted counts of objectives, with content as classified by expert panelists, indicate content emphasis for a given curriculum (Porter, 2006; Webb, 2007). McMaken and Porter (2012; see also Porter, 2006) indicate that while treating the count of items

by content category as a test content emphasis measure seems reasonable, the assumption that counts of objectives are an accurate measure of intended curricular emphasis could be problematic: in computing the SEC alignment index, each objective is weighted equally, "which we acknowledge may not reflect the intent of the [curriculum] authors but no other clear approach is apparent" (McMaken & Porter, p. 179). I first examine evidence of relationships between the coarse-grained curriculum proportions from SEC content analyses and concurrent measures of state curricular emphasis to address Research Question 1: ***Are counts of curriculum objectives a valid measure of curricular emphasis?***

If counts of objectives serve as a valid measure of curricular emphasis, instruction follows the formal written curriculum, and test item performance is sensitive to instruction, given sufficient examinee motivation and controlling for prior ability, test item performance should be positively related to the SEC's measure of proportional curricular emphasis for the corresponding objectives (see Mehrens & Phillips, 1987, for an analogous hypothesis regarding curriculum emphasis measures derived from textbook analyses). Finding of a substantively significant positive relationship between objective proportions and mean item-level achievement would suggest that all three of these conditions hold. Finding of a negative or null relationship would suggest that at least one of these conditions is false.

Models of curricular learning (e.g., Travers and Westbury, 1989) suggest that teachers' instructional emphasis reports should be a more proximal measure of state curricular emphasis than students' test item scores. If counts of objectives serve as a valid measure of curricular emphasis, and instruction follows the formal written curriculum, teacher self-reports of content coverage should be positively related to the SEC's measure of proportional curricular emphasis for broad content subcategories. Finding of a substantively significant positive relationship

would suggest that both conditions hold. Finding of a negative or null relationship would suggest that at least one of these conditions does not hold, or that the accuracy of the teacher survey data is poor.

The SEC alignment index assumes cognitive demand categories are nominal—only topic overlap by specific cognitive demand type is accumulated in the index. Other widely-used alignment methods characterize cognitive demand as an ordinal property of test items, and there is some evidence that these methods' cognitive complexity levels are positively correlated with item difficulty (Schneider, Huff, Egan, Gaines, & Ferrara, 2013). It has been asserted that the plausibility of this assumption is unlikely to affect "the overall substantive nature" of certain findings based on the index (Porter, Polikoff, & Smithson, 2009, p. 265). My next research question, Research Question 2, asks: ***Can the cognitive demand categories of the SEC content classification matrix be treated as partially ordered, rather than nominal as assumed by SEC alignment indices?*** The purpose of Question 2 is to check the appropriateness of the model underlying Question 1. If cognitive demand is best modeled as an ordinal property of test items, such that instruction requiring application of certain more-demanding cognitive processes related to a particular content topic also benefits students' ability to perform other less-demanding types of cognitive tasks related to the same topic (e.g., Ebel, 1956), models of the relationship between curricular emphasis measures (although not necessarily alignment) and achievement should account for the proportion of curricular content at or above a particular cognitive level. An affirmative conclusion regarding Question 2 would suggest that the Research Question 1 analyses should be repeated accounting for proportions of curriculum objectives at *or above* a particular item's cognitive level.

"The use of examinee response data to substantiate the apparent fit between a test and curriculum has been a long-running theme" in discussions of content validation (Crocker et al., 1989, p. 188, citing Gulliksen, 1950, Ebel, 1956, and others). Anderson (2002, p. 259) argued that "curriculum alignment enables us to understand the differences in the effects of schooling on student achievement" across, for example, courses or educational tracks. Test-curriculum alignment has been posited to affect school or state mean test scores, particularly in mathematics. Crocker et al. (1989, p. 188; see also Mehrens & Phillips, 1987) asserted that if various schools' math curricula do not match a particular test equally well, "there will be considerable variation in the schools' mean composite scores" due to variability in the degree of test-curriculum correspondence. It has similarly been reasoned that states with periodic mathematics assessment that is similar in content to the NAEP Mathematics test "might be expected to score higher [on NAEP] because of the alignment of curriculum with NAEP items" (Grissmer, Flanagan, Kawata, & Williamson, 2000, p. 112). If counts of objectives are found to be a reasonable measure of curricular emphasis, my final research question, Research Question 3, will attempt to provide empirical support for hypotheses that mean achievement test scores should increase with test-curriculum alignment: ***To what extent are item-level alignment measures related to achievement?***

To the extent that the content emphasis values underlying the SEC alignment index are meaningful, this analysis also responds to a methodological recommendation to incorporate opportunity-to-learn measures, as well as item feature indicators, into models of test item difficulty (Ferrara, Svetina, Skucha, & Davidson, 2011). Before detailing the methods that will be used to pursue these questions, I summarize the theoretical underpinnings of alignment indices, and previous empirical findings regarding their validity as evidence of test content

representativeness, and their relationship to student achievement test scores. My survey of the literature suggests that none of my three research questions have previously been addressed.

CHAPTER 2: LITERATURE REVIEW

Before detailing the methods that I will use to conduct the present study, I review the recent and historical research on alignment indices. I establish that alignment indices are a necessary type of validation evidence for particular types of educational tests. I describe elements of the data collection protocols implemented in various traditional and modern alignment methods. After outlining some similarities and differences among alignment methods frequently used for large-scale curricular achievement tests, I summarize the existing validation research that has been conducted to support the use of alignment indices in operational testing programs. While much of the previous research has focused on rating consistency among the subject-matter experts who are engaged to judge test-curriculum alignment, a few studies have reported on the expert raters' behavior. To situate my research questions in the literature, I emphasize open questions regarding the validity of alignment indices as quantitative measures of test-curriculum correspondence. By examining correlational relationships between Grade 4 mathematics curriculum content analysis data and other measures of state curricular emphasis, this study will contribute to validation of the SEC alignment index. Finally, I introduce previous work regarding the relationship between test-curriculum, instruction-curriculum, and instruction-test alignment and student test performance, particularly in mathematics. Although a positive relationship between test-curriculum alignment and students' test performance following instruction has been theorized to exist, previous empirical results have been mixed, with some apparently supporting, and others contravening this hypothesis. This study will investigate the influence of test-curriculum alignment on mathematics test item performance by Grade 8 students using recent data from ten US states, as further detailed in Chapter 3.

2.1 Alignment Indices as Validity Evidence for Achievement Test Score Interpretations

The amended ESEA requires that "assessments shall . . . be aligned with the State's challenging academic content and student academic achievement standards, and provide coherent information about student attainment of such standards" (No Child Left Behind Act of 2001). The US Department of Education has further ruled that evidence from formal test-curriculum alignment reviews of assessments of grade-level, modified or alternate curriculum objectives must be submitted prior to approval of state assessment systems. Guidance to state education agencies regarding documentation of their assessment systems specifies that the tests must:

> "Cover the full range of content specified in the State's academic content standards, meaning that all the standards are represented legitimately in the assessments; ***and***
>
> Measure both the content (what students know) and the process (what students can do) aspects of the academic content standards; ***and***
>
> Reflect the same degree and pattern of emphasis apparent in the academic content standards (e.g., if academic content standards place a lot of emphasis on operations then so should the assessments); ***and***
>
> Reflect the full range of cognitive complexity and level of difficulty of the concepts and processes described, and depth represented, in the State's academic content standards" (USED, 2004, p. 41, emphases in original text).

As well as being both a statutory and regulatory requirement for state assessment systems, test-curriculum alignment evidence is prescribed by theory for validation of the interpretation of achievement test scores as measures of curricular attainment.

2.1.1 State Achievement Test Scores Are Intended to Measure Attainment of Curriculum Goals

Validity is a quality of particular test score interpretations, not scores or instruments (Messick, 1989; Kane, 2006). Messick's work on validation has often been taken to imply that most, or all, educational test scores are intended to be interpreted as measuring examinees' standing on particular latent constructs, unobservable quantitative characteristics of persons (Peak, 1953). However, some educational test scores are not intended to have construct interpretations; rather they are intended as measures or predictors of observable traits, such as reading comprehension or complex mathematics problem solving (e.g., Kane, 2009; 2013; Millman & Greene, 1989)—the accepted meaning of the tests' item and total scores "derives from their action and outcome," not from posited relations among unobservable constructs, although such relations may exist (Guion, 1977, p. 6).

Some measurement professionals interpret the state curriculum objectives in a particular subject area as defining an achievement construct (e.g., Davis-Becker & Buckendahl, 2013), or view the objectives as representing only a subset of the content goals actually intended by a state (Koretz, 2008, p. 85). But others have concluded that typical contemporary state curriculum documents do not define achievement constructs. Haertel (1985) pointed out that educational outcomes tend to be defined "primarily in terms of their behavioral manifestations, and only secondarily in terms of cognitive processes" (p. 28; see also Guion, 1977). He suggested that subject area achievement outcomes are operationally defined by the objectives listed in state curriculum documents, which is consistent with the assumption of alignment indices that the objectives in a particular document "are intended to span the content of the goals . . . under which they fall" (Webb, 2007, p. 9; see also Haertel, 1985, p. 28). Compared to broader item domains evoked by potential alternative construct characterizations of achievement, state

curriculum documents tend to suggest relatively unique item domain specifications that may be most appropriate for measuring observable traits (Haertel, 1985). Furthermore, to compose the test specifications for state achievement tests, test developers typically do not refer to any broader complex achievement construct (Ferrara & Duncan, 2011, pp. 143–144), implying that achievement scores should be considered measures of an observable trait—performance on tasks with academic content—possibly measured with systematic or random error.

Although legislators may wish to ascertain general academic performance,—that is, they may perceive state tests' target domain to be, for example, academic mathematics achievement—because tests are usually developed and assembled under the stricture that each item match at least one objective in the state curriculum, the scores' potential universe of generalization (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) effectively covers, at most, listed objectives and their components, but not similar objectives that could potentially have been included in the curriculum. In practice, the target domain for a state achievement test might more reasonably be considered to be the union of the potential task sets corresponding to each curriculum objective, including performance tasks accomplished under a variety of administration conditions. Because, in many states, some objectives are consistently excluded from the item development process for practical reasons (Ferrara & Duncan, 2011), the target domain may be further narrowed to include only potential items related to "testable" curriculum objectives (Martineau et al., 2007). However, making inferences about expected student performance on a stated list of curriculum objectives when tasks are presented in novel or real-life contexts may still be an ambitious goal. The universe of generalization, the description of the behavior types to which test scores can reasonably be generalized (Cronbach et al., 1972), for state achievement tests may

be limited to performance on tasks from the target domain that are similar to those called for during testing, e.g., constrained-response items, standardized administration conditions.

If curricular domain definition is adequate, and items approximate random sampling from the task universe, a valid limited descriptive inference from the test scores immediately follows: an examinee's estimated true score is the mean proportion of items in the item universe that the examinee would be expected to answer correctly on that measurement occasion under the given administration conditions (Linn, 1980). While government entities may wish to extrapolate to performance on a broader domain—under additional administration conditions or on tasks that are not feasible to incorporate in large-scale on-demand testing, these types of inferences generally require evidence beyond that which can be gleaned from statistical scoring models (Kane, 2013). If inferences about classroom or school performance, rather than or in addition to individual student performance, are desired, minimally from an item content perspective, the aggregate item set administered at that level must approximate random sampling from the item universe.

Currently, state achievement test scores are intended primarily to produce mean proficiency classifications at the school or classroom level, and secondarily to generate individual proficiency classification estimates. Schools' mean proficiency levels are used as one component of states' school ranking system, ratings from which are disseminated to the public and used to identify underperforming schools. Recently, the federal government has also incentivized states to use classroom-level achievement estimates, either mean scores or proficiency levels, to evaluate teacher performance (e.g., Notice of Final Priorities for Race to the Top Fund, 2009).

2.1.2 Alignment Evidence is Necessary for Validation of Intended Achievement Test Score Interpretations

State achievement tests should "either representatively sample or comprehensively measure" the assessable curriculum objectives in the same proportions as they appear in the complete set of objectives (Martineau et al., 2007, p. 30). Content evidence is necessary for validation of any score interpretation because such evidence connects a test to the trait it is intended to measure, whether conceived as a construct or an observed variable (Yalow & Popham, 1983). Yalow and Popham (1983) argued that it was barely conceivable, and certainly undesirable, to use a test that poorly corresponded to a particular target domain to draw inferences about examinee performance in that domain. Information about test content is highly relevant to the meaning that test users can attribute to scores, and becomes more salient to validation as a test's target domain is increasingly "rooted in behavior with a generally accepted meaning" (Guion, 1977, p. 6; see also Kane, 2009). Whether test scores are interpreted as measures of a construct or an observable trait, evidence that test content is representative of an intended curricular domain is required for validation of score interpretations that refer to the curriculum (Crocker & Algina, 1986). If it can be shown that a test constitutes a representative sample from the content domain of interest, an examinee's score on the test can be expected to reflect how the examinee will perform on domain tasks (Yalow & Popham, 1983).

Kane's (2006, 2013) argument-based approach to validation suggests evaluating the plausibility of the claims and assumptions justifying each distinct test score interpretation. He describes major types of inferences that would be contained in most score interpretation or use arguments, including (a) scoring inferences, in which observed scores (e.g., weighted sum scores) are inferred from observed performance based on particular warrants and their backing, (b) generalization inferences, in which universe scores (e.g., item response modeling "theta"

ability estimates) are inferred from observed scores, typically based on a statistical model, and (c) extrapolation inferences, in which domain scores (e.g., standards-based classroom grades predicted from curricular achievement test scores [Welsh, D'Agostino, & Kaniskan, 2013]), which represent some performance conditions that were not sampled or observed, are inferred from universe scores. Because generalization inferences rely on sampling theory, the item sample that composes each test must be argued to represent the item universe, although it rarely could be argued to be random. Alignment indices provide evidence regarding the extent to which a test's item set is a representative sample from the relevant curricular domain (McMaken & Porter, 2012), as would be expected under simple random sampling. To the extent that these indices are accurate measures of test content representativeness, they serve as an important warrant for claims in generalization inferences, and extrapolation inferences that build on these generalization inferences.

The validity of a particular score interpretation can serve as a warrant for score use arguments that involve that interpretation (Kane, 2013). Although evidence of items' sampling adequacy may suggest particular test score interpretations, it does not suggest any particular test score uses, and can support particular uses only indirectly as those uses, often decisions based on the scores (Kane, 2013), involve specific score interpretations. For example, test content that is not taught because it does not appear in the curriculum could influence test scores, and "if one infers that instruction was faulty . . . teachers could be inappropriately blamed" (Mehrens & Phillips, 1986, p. 186). Since alignment index values, and alignment judgments more generally, do not imply any particular uses of state achievement test scores (e.g., diploma conferral, teacher evaluation), in this paper I focus on the indices' contribution to validation of score interpretations.

2.2 Measuring Test-Curriculum Correspondence: Traditional and Modern Alignment Methods

Although some traditional methods for collecting evidence of test content relevance or representativeness include measures and evaluation criteria related to alignment, as I have defined it, the development of alignment procedures during the late 1990s and early 2000s (e.g., Rothman, 2003; Webb, 1997) occurred largely without reference to research on these existing methods (Martone & Sireci, 2009). Early theorists defined alignment as "a function of how well the test content matches the curriculum content domain" (Guion, 1977, p. 7), and proposed assessing it by classifying items into "broad areas" of subject matter and behavioral performance type (Ebel, 1956, p. 275). Subsequently, numerous quantitative techniques for collecting content-related evidence, all of which rely on matching test items to some representation of the target domain, whether, for example, a broad domain definition, a test specifications document apportioning items to various content categories, or a curriculum document detailing specific behavioral objectives in a content area, were developed (Crocker & Algina, 1986).

2.2.1 Traditional Evidence of Test-to-Specifications Alignment

Traditional methods for collecting validity evidence based on test content focus on three distinct types of match: item-objective congruence, test-instruction congruence, and test-test specifications congruence. Indices of item-objective congruence are computed from the proportion of raters matching each item to its item-writer-intended objective (Sireci, 1998) and the quality of the match (Rovinelli and Hambleton, 1978, cited in Hambleton, 1980; Crocker & Algina, 1986, p. 221; Turner & Carlson, 2003; Sireci & Geisinger, 1992). For any index of item-objective congruence, Crocker, Llabre, and Miller (1988, p. 288) suggested taking the average of mean ratings over items as a measure of test-curriculum match. A serious limitation of these

congruence indices is that the matching process tends to be “extremely” time consuming because judges must compare each item to every objective (Crocker et al., 1989, p. 185). Also, the potential magnitudes of these indices are influenced by both the number of raters and the number of objectives, so no fixed criterion for an acceptable value of each item index can be set; situation-specific criterion must be utilized (Crocker et al., 1989).

Jones and Szatrowski (1983) proposed three content validity criterion alternatives, all of which were based on the assumption that the validity of test scores is related to the proportion of examinees in the population who have received instruction relevant to correctly answering each test item. The most complex of their proposed criteria required a minimum level of population exposure to relevant instruction, which was estimated from teacher surveys, covering a user-determined minimum proportions of items within each major subtopic appearing on the test. Klein and Kosecoff (1975) suggested use of the correlation between the importance weighting of each curricular objective and the number of items measuring each objective as an index of test-curriculum match. However, the magnitude of this correlation index is affected by variance in the numbers, or importance weightings, of items corresponding to each objective; the correlation would tend to be reduced as objectives were weighted as equally important, or represented by equal numbers of items (Crocker et al., 1989).

A simple index proposed by Rovinelli and Hambleton (1978, cited in Hambleton, 1980) was computed from several raters’ item-objective match data. A chi-square independence test on the item-by-objective contingency table composed of counts of raters matching each given item with particular objectives has a straightforward interpretation as a test of whether the item set is significantly associated with the curriculum objectives, perhaps a minimal requirement for establishing test content relevance.

2.2.2 Comparison of Traditional and Modern Alignment Methods

Both traditional and modern alignment methods focus on items as the appropriate unit of analysis from the test instrument, and rely on item matching or rating with reference to a defined target domain (Crocker et al., 1989; D'Agostino et al., 2008). Although not a methodological requirement, most applications of traditional content methods have matched test items to the broad content and cognitive demand categories comprising the test specifications (e.g., Sireci, 1998, p. 300; Schmeiser & Welch, 2006; Crocker & Algina, 1986, p. 219) rather than to the more specific content represented in detailed curriculum objectives (Martone & Sireci, 2009, p. 1336). If content matching is conducted with reference to the test specifications, the matching procedure only addresses correspondence of the test to those curriculum objectives that are covered by the test specifications, and does so only indirectly. Recent alignment methods specifically recommend analyzing the most detailed level of behavioral performances listed in a curriculum document (Porter, 2002; Webb, 2007), and published applications have consistently followed this instruction (e.g., Webb, Alt, Ely, Cormier, & Vesperman, 2005; Roach, McGrath, Wixson, & Talapatra, 2010). However, validation studies using traditional methods have also sometimes conducted content matching with respect to a formal curriculum document (e.g., Klein & Kosecoff, 1975). Traditional content-matching methods that focus on quantifying test content representativeness, such as the indices of item-objective congruence described previously, can be considered alignment methods under our definition, so previous research on the underpinnings and limitations of these methods, although very limited (Crocker et al., 1989), is relevant to hypotheses about alignment index performance.

2.3 Connecting Test Items to Curriculum Objectives

Procedures used to link test item and curriculum objective content can differ along several dimensions, including the types of task features that are considered relevant to judging alignment, the reporting of error variability among panelists, whether an intermediate content classification table is used, and whether connecting items to objectives involves binary matching, ordinal rating, or both. Alignment procedures can be classified into two general categories: (a) methods that directly match items to objectives (e.g., Frisbie, 2003), and (b) methods that estimate the proportion of items that match each objective (e.g., SEC; Davis-Becker & Buckendahl, 2013), which indirectly match items to objectives by directly matching both to a set of test content categories. Anderson (2002) recommended use of a generic content taxonomy table to conduct alignment in any subject area. She argued that mapping educational document content to a generic table should be preferred over directly matching document units (e.g., test tasks, curriculum objectives) because the act of classifying content items "focuses quite directly on student learning," while lessening the tendency of political or personal implications of the results to influence judges' ratings (p. 258).

Alignment procedures also vary in the types of item and objective features that are considered in determining alignment. In the simplest methods, alignment judgments are based only on the extent of correspondence between items' and objectives' content topics, making it "more likely that a match will be found" than if judgments also consider correspondence on cognitive demand or other task features (Bhola et al., 2003, p. 24). Because USED (2004) requires alignment evaluations of states' Title I accountability test systems to account, minimally, for their correspondence with curricular content and cognitive demand, the most frequently used alignment methods in K–12 achievement testing all classify items by content and

cognitive demand—low complexity alignment models (Bhola et al., 2003) will not suffice to meet the requirement.

Well-documented alignment procedures typically include scripted or written instructions that detail the method all raters should use to identify task topics and cognitive demand, for instance by focusing on tasks' verbs and nouns (Anderson, 2002; D'Agostino et al., 2008), discouraging raters from developing their own idiosyncratic rules (Webb, 1999). The instructions typically also include guidance regarding matching items to multiple objectives, but this tends be general, rather than situation-specific, and not necessarily consistent with instructions given to item writers (Davis-Becker & Buckendahl, 2013). Development of the methods that combine individual judges' ratings to produce an overall alignment index has tended to emphasize the importance of maintaining independence among raters. Some applications of the SEC procedure have encouraged judges to discuss any problems or questions after initial coding before giving a final coding of each objective (McMaken & Porter, 2012). The Webb (1999) procedure requires judges to reach consensus on the cognitive demand codes for objectives, but directs them to conduct item-objective matching and rate item cognitive demand independently.

Most existing alignment procedures can be classified as implementing rating, in which expert judges rate the strength of correspondence between each test item and a pre-assigned objective, typically from the test specifications, or matching, in which judges determine which objective or objectives from a list most closely corresponds to each test item (D'Agostino et al., 2008). D'Agostino et al. (2008) randomly assigned 49 subject-matter experts either to match high school mathematics achievement test items from Arizona to curriculum objectives or to rate the strength of item-objective links. Raters judged content, cognitive demand and overall consistency between item and objective pairs using three-point scale, while matchers matched

each item to up to three corresponding objectives. The authors found that itemwise alignment decisions made using the rating and matching methods agreed moderately, with a correlation of .59 between average alignment indices for each item. Rating was more time efficient than matching, requiring about 25% less time, but, the authors believed, was more likely than matching to encourage acquiescence or, more generally, rater leniency, particularly if objectives were so broad that they could plausibly be measured by a wide range of items.

2.4 Defining Cognitive Demand Categories

Snow and Lohman (1989) proposed that examinees' observed item performance be interpreted as samples of their cognitive processes, which were inherently unstable, rather than as signals of their standing on a well-defined, although unobservable, latent trait. They held that the cognitive processes used to complete a task would vary among, and possibly within, examinees on a single measurement occasion, depending on examinees' physical and social situations, as well as their perceptions of the task's components. However, others assert that assuming common instruction of the examinees, it may be possible to average over their responses, treating the cognitive process required for correct response to a test item as a fixed property of the item in a given population. Snow (1994) allowed that as people share a common learning history during socialization, schooling, or job training, "common patterns of ability will be seen to develop," although he believed that classifications based on shared instructional experience "may leave out more important information about persons. . . than they capture" (p. 15). Mislevy (2009) argued that to the extent examinees' context includes common instruction and life experiences "students' propensities for actions in…task situations *can* be said to exist" (p. 100; emphasis in original), producing response patterns that can be modeled. Unfortunately, often, not even instructional histories are known, so cognitive processes used by "even a majority

of the test takers" must be inferred by test developers and reviewers from highly distal evidence (Schmeiser & Welch, 2006, p. 316).

Modern alignment methods all rely on classification of test items and curriculum objectives into mutually exclusive categories of cognitive demand or complexity. Cognitive complexity coding schemes may reflect item linguistic features, many of which, in the case of mathematics achievement items, would be considered sources of nuisance response variability unrelated to the trait of interest, item structural features, which are central to measuring the trait (Lepik, 1990), or both. Early content validation methods categorized test items based on their content and the type of performance they required of examinees (Ebel, 1956). Ebel (1956) explicitly stated that the performance categories did not assume use of any particular cognitive processes by examinees, only types of observable performance. Some extant cognitive demand classification schemes encode item features, requiring few assumptions about cognition (e.g., Lepik, 1990; Schneider et al., 2013). However, consistent with modern curriculum development efforts' reliance on taxonomies of cognitive performance to categorize statements of each objective, modern alignment methods require judges to make inferences about examinees' cognitive processing.

Item cognitive demand ratings can be defined as "the baseline level of cognitive processing required to provide a correct response" (Wyse & Viger, 2011, p. 188), or as intended to reflect the solution process most examinees, or average examinees, use to solve an item (Schmeiser & Welch, 2006). Cognitive demand is invariant to changes in an item's context, and to modifications affecting only item content, but not the solution process (Wyse & Viger, 2011). Item cognitive demand is distinct from the concept of item difficulty, although ordered cognitive complexity ratings would be expected to have a systematic relationship with observed item

difficulty, the average probability of correct response (Embretson & Daniel, 2008; Gorin, 2006; Wyse & Viger, 2011), and tend to be related to observed difficulty in practice (Martineau et al., 2007). Cognitive demand may be viewed as a property of test items, not jointly of test item-examinee population combinations, so that the cognitive demand of a test item does not necessarily change across examinee populations and is independent of the specific curriculum to which each examinee has been exposed, but this perspective also requires the assumption that all test takers are familiar with the general approach to each task (NAGB, 2006). Other interpretations suggest that true item cognitive demand is tied to particular examinee populations, who may tend to reach correct solutions in distinct ways (Roach, Niebling, & Kurz, 2008), depending on their instructional background (Embretson & Daniel, 2008; Schmeiser & Welch, 2006). For example, in states where particular well-known number sequences (e.g., the Fibonacci sequence) are part of the curriculum, related test items may tend to require a lower level of cognitive demand from students than in states where these sequences are not explicitly covered, and students will have to reason to reach the solution (Sanford & Fabrizio, 1999).

Even when most examinees follow the same instructional sequence, item cognitive demand ratings will depend on the extent to which a given classification scheme accounts for specific characteristics of the examinee population. Consider, for example, a test item that requires examinees to recall an obscure historical fact, which was an element of all examinees' instruction, but was not highlighted. If the classification scheme focuses raters' attention on the generic type of cognitive process, the verb—recall—ratings are likely to be different (in this case, lower, if an ordinal scheme is used) than if the classification scheme directs raters to consider the demand of the specific cognitive process typically activated during examinee-test item interactions in this population.

Ebel (1956) indicated that categories for different types of behavioral performance should be considered at least partially ordered by degree of difficulty. Taxonomies of cognitive performance (e.g., Bloom, 1956) similarly prescribe ordered categories. There is some evidence that test items of varying formats require different types of cognitive performance, and that these performance types can be ordered by complexity (Martinez, 1999). Among alignment methods, the Webb (1997) and Achieve (Resnick et al., 2004) methods represent item cognitive demand as a set of ordered categories, while the SEC method (Porter, 2002) utilizes nominal categories for cognitive demand. While each coding scheme may capture unique elements of a hypothesized item response process, many frameworks' demand category definitions describe similar levels or types of processing; these commonalities may be reflected in relationships between the ratings from various classification schemes. For instance, when two raters applied several different coding schemes including reading load, NAEP mathematical complexity, and Webb depth-of-knowledge to characterize math item cognitive demand, their depth-of-knowledge ratings were significantly positively correlated with their mathematical complexity ratings for both Grade 4 and 8 items, and with their reading load ratings for Grade 4 items (Schneider et al., 2013). Similarly, when considered pairwise, some modern alignment methods' cognitive demand categories, whether characterized as ordinal or nominal, appear to overlap in meaning, but these apparent relationships have not been substantiated empirically.

The use of Bloom's (1956) taxonomy of cognitive levels to guide item and test development has been widely criticized (see, e.g., Hattie, Jaeger, & Bond, 1999, p. 405, for a summary), and the need for an empirically-supported taxonomy of cognitive behaviors to guide item writing has been pointed out (Haladyna et al., 2002; Schmeiser & Welch, 2006). Similar questions can be raised about the cognitive demand categories applied during alignment

procedures. Many cognitive complexity item coding schemes, including commonly-used schemes like Webb's (1997) depth-of-knowledge scale and the NAEP mathematical complexity scale, have little or no empirical support (Ferrara, Svetina, Skucha, & Davidson, 2011). For most coding schemes, there is limited evidence that cognitive complexity ratings, and corresponding rating category descriptors, accurately portray aspects of a typical examinee's item response process (Embretson & Daniel, 2008; Webb, 2007). In one empirical study (Schneider et al., 2013), the poor prediction of item difficulty provided by five different cognitive complexity rating schemes was partially attributed to the wide distributions of observed item difficulties in the lowest categories of all the rating schemes, suggesting that some important distinctions among item features were not captured by the descriptors for the lowest rating categories, or that the specific cognitive processes applied by the examinee population did not correspond to the category descriptors. A further caution is that even if the rating category definitions are sound, subject-matter experts are seldom able to accurately predict the cognitive processes used by examinees to solve achievement test items (Ferrara et al., 2004).

2.5 Establishing Alignment Criteria

Alignment results "depend critically" on the definitions of the criteria utilized (Bhola et al., 2003, p. 24). Even if two alignment procedures utilized the same model (e.g., the Webb [1999] "Content" model), with component aspects both labeled and defined identically, alignment decisions would clearly depend on the stringency of the criteria applied to values on each component, or to the overall score. Webb (2007) suggests that if a consistent alignment model is utilized across studies, it may be possible to devise experience-based criteria for desirable alignment index magnitudes. Alternatively, given alignment index values from pairings of many states' curricula and assessments, it may be possible to make normative

judgments about alignment magnitude. However, no alignment indices have cutoff criteria that have been devised based on empirical research, or are widely agreed upon (Davis-Becker & Buckendahl, 2013). Summary alignment reports generally reflect the conflict between beliefs that alignment cannot be meaningfully quantified (e.g., Beck, 2007), and that alignment indices possess approximately interval properties and contain particular scale points that can be given meaningful interpretation (e.g., Webb, 2007). Criteria for "acceptable" overall alignment index values rely on assumptions that may be difficult to justify in some testing contexts (Webb, 2007). While reports often present alignment index values, giving them either absolute (e.g., Webb, 2007) or relative (e.g., Polikoff, 2012a) interpretations, they tend to be situated in a broader evaluative narrative that attends to item balance across particular content types, items flagged as irrelevant, and contextual issues, such as the test purpose, and level of resources available (Webb, 1997) for test and curriculum development.

A minimum criterion for overall alignment could be that an alignment index value is significantly greater than would be expected due to chance agreement between the ratings of test items and curriculum objectives covering a given subject matter (Fulmer, 2011). Fulmer (2011) demonstrated a method for estimating critical values of an SEC-type alignment index, computed from proportions in a content taxonomy table, for various given statistical significance levels, table sizes, and numbers of items and objectives. He also verified through simulation that the estimated mean index values expected by chance, if judges coded both tasks and objectives randomly without regard for their content, would tend to decrease with increasing table size for a fixed number of coded items or objectives, and increase with the number of coded items or objectives for a fixed table size. Alignment index values expected by chance would also increase

with the number of raters, with the number of items or objectives coded to multiple table cells, and with decreasing rater agreement (Polikoff & Fulmer, 2013).

Alignment criteria focused on individual tasks, rather than entire documents, have also been proposed. La Marca et al. (2000) argued that, minimally, all test items should be relevant to the curricular domain. Gulliksen (1950) asserted that educational tests should not contain items that required novel applications of learned content unless examinees have had previous practice with such new applications, because otherwise the tests would likely be perceived as unfair by examinees, which might negatively affect their attitudes toward test-taking, or future learning in the content area. Webb (1997) concurred with Gulliksen: "expectations and assessments are aligned if what is elicited from students on the assessments is as demanding cognitively as what students are expected to know and do" (p. 15). However, the issue of how to deal with planned item- or objective-level misalignment (e.g., Woolard, 2007, p. 11) in computing alignment indices has not been resolved. Some objectives that cannot be feasibly or efficiently tested by large-scale assessment are consistently omitted from states' achievement test specifications. Additionally, items appropriate for students at lower or higher grade levels may be included in a test (e.g., Webb, 1999) to facilitate score scaling. Published alignment studies have typically included all objectives and items in alignment index calculations, although have occasionally, at the request of particular states, recomputed the indices using only testable objectives (see MECG, 2010).

2.6 The Validity of Alignment Indices as Evidence of Test Content Representativeness: Previous Empirical Findings

Previous research on alignment results' validity as measures of test content representativeness has focused on an issue common to all judgmental alignment procedures: the

quality of the ratings generated by alignment panels. In evaluating test-curriculum correspondence, "the number of judges used, their competence, and the process they use in evaluating the test...and the conscientiousness with which they undertook the task of evaluation" help to determine the quality of their judgments (Ebel, 1956, p. 278). Recognizing that the value of any alignment data collected hinges on judges' expertise and adherence to a consistent rating process, published alignment methods provide guidelines regarding assembling and training the panel, which have been modified over time based on empirical findings.

Judges should be subject-matter experts who are familiar with the abilities of students in the target population, and may include university faculty, state department of education employees or consultants, graduate students with advanced degrees in the content area, or classroom teachers (Porter et al., 2008; Webb, 2007). The most qualified judges of how well test content corresponds to a particular curriculum content domain are those with the greatest degree of "knowledge of the curriculum in a specific school system," rather than those with "abstract, generalized curriculum ideas" (Guion, 1977, p. 7). Alignment panelists must possess both content area expertise and knowledge regarding typical abilities in the student population to be assessed (Davis-Becker & Buckendahl, 2013). Ideally, alignment panelists should have knowledge of the specific curriculum document used in the matching procedure (La Marca et al., 2000). Repeated use of the same panelists may improve comparability of alignment results across different test-curriculum combinations; however, in this situation any rater or panel bias could introduce systematic error into a set of alignment results. Representative sampling of content judges is a fairness issue (Guion, 1977); alignment panelists should be representative of stakeholders in the assessment results (Davis-Becker & Buckendahl, 2013).

The minimum number of judges recommended by each alignment method varies, but is often lower than the minimum number recommended for standard-setting panels (e.g., 15–20; Hambleton, Pitoniak, & Copella, 2012). Webb (1997, 2007) has recommended use of between 3 and 8 subject-matter experts for an alignment panel. SEC content analyses are typically conducted by between 3 and 5 raters, although sometimes as few as two raters have participated (Porter et al., 2008). The Achieve method (Resnick et al., 2004, p. 8) requires at least 6 judges. As during standard-setting studies, the "personality, skill, biases, perspectives, and personnel management abilities" of the facilitator are also important variables mediating the quality of data collected from an alignment panel (Beck, 2007, p. 131).

Published alignment methods provide written materials to guide consistent implementation of instruction by facilitators. Variation in phrasing of instructions to panelists may affect their ratings of each item (Poggio et al., 1986; Bhola et al., 2003). Webb (1999) reported that when judges were given little initial guidance in interpreting rating categories, individual reviewers and groups of reviewers developed their own decision rules for coding. If applied consistently by several raters, self-developed coding schemes could lead to systematic error in measuring alignment as compared to under the (in this case, unstated) intended coding rules. Requiring reviewer training on a rubric as the first step of alignment process encourages reviewers to hold common definitions, for example, of cognitive complexity categories (La Marca, 2001). While developers of alignment methodologies have emphasized the importance of allotting sufficient time for training to permit panelists to practice and thoroughly understand the coding process (e.g., Porter, 2002; Webb, 2007), the time actually expended on training varied markedly across early implementations of the various alignment procedures (Rothman, 2003) depending on the resources of the sponsoring organization. As in standard-setting studies, the

amount of time allotted for training, practice, and document content analysis, is likely to influence panelists' understanding of a given alignment process and confidence in their judgments (Martone, 2007).

Sireci and Geisinger (1992) argued that content raters' judgments should be independent of information regarding item writers' intent, and even of pre-specified content categories. Given an objective list, Sireci and Geisinger (1992) expressed concern that raters might tend to match items to the "closest" objective, rather than considering potential alternative objectives, not listed, that might more closely correspond to an item. The provided objective list is likely to influence subject-matter experts' perceptions of what each item is measuring (Sireci, 1998, p. 303). "By informing the [subject-matter experts] of what the test is supposed to measure, item congruence . . . ratings can be influenced by response sets such as social desirability and guessing" (p. 303), possibly inflating item-objective congruence index values. To avoid inducing rater response sets, which could bias item ratings, by provision of a content categories list from the test specifications, Sireci and Geisinger (1992) developed an item-similarity matching method based on multidimensional scaling. However, Martineau et al. (2007) recommended that upon identifying mismatched items, alignment panelists should be advised of item writers' intent in writing the item. Advisement of item writer intent should increase the precision, but also possibly the bias, of content ratings.

Sireci (1998) listed several possible threats to the validity of item-objective congruence measures, all of which concern the quality of judges' ratings or matches: (a) poor reliability of ratings due to an insufficiently large rater sample, rater fatigue, or the inherent complexity of the rating task, (b) poor comprehension of the rating task by judges, and (c) bias caused by rater response sets induced by provision of a fixed objectives list. Although no published alignment

procedure is supported by a systematic program of research testing for the presence of these confounds, empirical studies of modern alignment indices have addressed the first two threats, meanwhile uncovering suggestive evidence of characteristics associated with rater bias.

2.6.1 Alignment Index Reliability

Crocker et al. (1988) proposed computing a generalizability coefficient at the end of content-based item analysis, or in a preliminary generalizability study, to check that the number of raters utilized produces item-objective correspondence index values that are adequately replicable in repeated independent sampling of rater panels. They used an analysis-of-variance model to decompose the variance of item ratings, attributing different portions of the total variance to various specified possible sources (e.g., raters, item content), and computed generalizability coefficients, of which traditional alpha reliability coefficients are a special case, to assess the variability in test scores attributable to particular random features of a measurement procedure, in this case, raters, for various potential numbers of panelists. They noted that the projected generalizability coefficients apply only to the given test domain specification and rater population, since the number of raters needed to produce reliable item-objective correspondence indices is likely to depend on the breadth of a test's target domain, as well as the specificity with which objectives are written.

Porter et al. (2008) measured the magnitude of rater effects, cellwise, on the matrices of fine-grained content emphasis proportions from SEC content analyses of English language arts and mathematics achievement tests and curriculum documents from two states for grade levels 3, 6, and 9–12. These matrices of average proportions underlie the alignment indices computed from SEC data. For all state-subject-grade-document type combinations, the value of the generalizability coefficient prediction approached an asymptote above .9 as the number of raters

reached approximately eight or nine (p. 4). In one state, rater generalizability was lower in English than in math for all grade levels and document types. Otherwise, projected generalizability estimates for a given number of raters were fairly consistent regardless of grade level, subject matter, and whether a test or curriculum document was analyzed. The authors concluded that future SEC alignment review procedures should recruit at least five raters, although generalizability coefficients were mostly acceptable, exceeding .70, with four raters. Herman, Webb, and Zuniga (2007) also computed an index of dependability-type generalizability coefficient for the Webb depth-of-knowledge alignment index calculated from 20 judges' content ratings of California's high school mathematics achievement test and a statement of mathematics competencies expected of freshmen entering University of California system institutions, estimating a .90 dependability coefficient. While the number of panelists in that study was much higher than is typical in alignment studies of state achievement tests, existing research, overall, seems to indicate that adequately reliable alignment index values could be obtained by recruiting more panelists than were recommended when these alignment methods were initially developed, perhaps between about 8 and 15 panelists.

2.6.2 Rater Agreement

When a particular item is matched to different objectives by reviewers, these judgments suggests panelists attribute "diverse" meanings to task statements, particularly with regard to their content and cognitive demand (Herman et al., 2007, p. 122), and may represent legitimate differences of opinion that would be expected when applying an inherently judgmental procedure to possibly complex test items (Rothman, 2003; Webb, 1999). However, disagreement could indicate a problem with the clarity of task content, with panelists' interpretation of the alignment matching criteria (Davis-Becker & Buckendahl, 2013), or with panelists' decoding of task

content (D'Agostino et al., 2008). Thus, sources of substantial disagreement among alignment panelists should always be investigated (Davis-Becker & Buckendahl, 2013).

In some situations, disagreement may be attributable to characteristics of the documents analyzed, or to the alignment procedures themselves. When a single curriculum document includes descriptions of intended student performance at multiple levels of specificity (e.g., detailed objectives are subsumed under broad content "strands" or subcategories), agreement will tend to be higher as items are matched to broader (e.g., compute basic operations), rather than narrower (e.g., subtract three-digit whole numbers), performance descriptors, simply "because there are fewer opportunities for disagreement" (Davis-Becker & Buckendahl, 2013, p. 27). Similarly, for indices that measure test-curriculum overlap based on the proportional match of items and objectives in a content matrix, such as the SEC alignment index, if the matrix is very large, requiring reviewers to simultaneously attend to many content categories during matching, agreement may tend to be much lower than if the matrix has only a few cells describing broad content categories (Mehrens & Phillips, 1986, p. 186). In other cases, lack of agreement among panelists "might be due to characteristics or behavior of the raters themselves, including insufficient training on the rating process, insufficient depth of understanding of the standards, lack of content knowledge, inappropriate use of secondary objectives, fatigue, and coding errors (mistakes in writing down the appropriate objective number)" (Webb et al., 2007, p. 25). If alignment indices reflect underlying disagreement that suggests some raters may be seriously misinterpreting task statements, or some task statements are too vague to interpret with any confidence, the alignment indices' may become "a function of who does the rating rather than a function of a test item's content and cognitive demand" (Webb, Herman, & Webb, 2007, p. 25), compromising the indices validity as measures of content representativeness.

To examine rater agreement during alignment procedures, Herman et al. (2007) used ratings from a panel of 20 judges who rated high school mathematics behavioral objectives and test items from California, as detailed previously. Training was completed on the same day the ratings were collected, and followed established recommendations. Recognizing that most alignment studies rely on considerably smaller numbers of judges, consistent with recommendations in the literature (e.g., Porter, 2002; Webb, 2007), they simulated a more realistic quantity of alignment data by drawing ratings of all possible 6-judge subsets, each composed of three high school math teachers and three university faculty, from the 20 judges. Depending on the judge subset selected, proportion agreement reached at least .65 (a criterion set by the authors; they recommended that for an item to be included in an alignment index computation, at least 65% of raters must match it to the same objective, or assign it the same cognitive demand rating) on the general content category for most of the 42 items, between about 75% and 100%, depending on the judge subset selected, but agreement on the specific content topic, number of topics, and cognitive demand measured by each test item varied more widely, between 50% and 100% depending on the particular panel assembled. If multiple item features, e.g., topic and cognitive demand, were considered simultaneously, proportion agreement among the 6 judges in each subset tended to be even lower, as anticipated. The authors concluded that with only 6 panelists, agreement about item-objective content match was limited.

Webb, Alt, Ely, Cormier, and Vesperman (2005) analyzed rater agreement in 34 selected alignment studies that used the Web Alignment Tool, an online implementation of the Webb method. They found that rater agreement on item cognitive demand levels, measured by intraclass correlation (ICC) and an average pairwise agreement statistic, was usually determined to be acceptable, with ICCs greater than 0.7 and pairwise agreement greater than 0.6. Four of the

alignment studies, two of which had low variability in mean assigned cognitive demand levels among items, and two of which had only three raters, were judged to have unacceptably low interrater agreement. Rater agreement on item cognitive demand levels tended to be higher for lower grades' curriculum-assessment pairs, which tended to include more items that could be assigned to the lowest demand category with high certainty (p. 18). Rater agreement in item-goal matching, measured by pairwise agreement, was usually acceptable. However, rater agreement in matching items to specific objectives under each goal, again measured by pairwise agreement, was less than .5 in nearly two-thirds of the studies, including several in which eight or nine raters participated. Rater agreement in item-objective matching was lowest for studies utilizing curriculum documents with the largest number of objectives. Slight improvements in agreement over time were attributed to improvement in the training materials.

The Webb and SEC alignment methods do not require interrater agreement for item-to-objective matches or task-to-content taxonomy table classifications, respectively, even at the level of broad content category, and any disagreement tends to be masked by their indices, which rely on averaging (Davis-Becker & Buckendahl, 2013). Many traditional alignment methods similarly average over all ratings, regardless of the extent to which they agree. Martone (2007) reported that in one alignment study, many items were counted as a "match" to multiple specific objectives, some of which refined different broad content goals. Because failure to resolve or account for this disagreement in any way may be problematic for use of itemwise alignment results in test revision (Davis-Becker & Buckendahl, 2013), and for meaningful interpretation of alignment indices, removing some items' or raters' data from alignment computations has been suggested. Herman et al. (2007) proposed that when raters do not reach some prespecified level of agreement (they recommended 65%) in matching particular items, those items should be

excluded from alignment index computations. Using data from three previous Webb-type alignment studies that compared (a) Michigan curriculum objectives to state achievement test forms, (b) Tennessee curriculum objectives to state achievement test forms, and (c) California's high school mathematics exit examination to math standards expected of entering freshmen by the University of California system, Webb et al. (2007) found that when they recalculated the four Webb alignment indices using only items for which raters reached a minimum level of agreement (either a bare majority, or a clear majority), and applied Webb's (1999) alignment criteria, conclusions about each aspect of alignment often differed from the original conclusions. Porter et al. (2008) reported that in computing generalizability coefficients for two states' item and objective content classification tables, the results for one state included two aberrant sets of ratings at different grade levels. Generalizability estimates improved when these ratings were omitted, perhaps implying that any such aberrant judges' ratings should also be excluded from alignment calculations.

The existing research suggests that the amount of disagreement being averaged over to compute alignment indices sometimes has been high enough to warrant concern about the indices' accuracy. Transparent alignment review results report the level of rater agreement obtained, flagging any discrepant raters or items. Techniques intended to address lack of agreement in alignment ratings include enlisting larger numbers of reviewers, averaging results among reviewers, and improving training (e.g., Webb et al., 2007). Presenting corrected alignment indices that exclude data from particular raters or items if evidence suggests problems with rater comprehension of certain items, or systematic rater bias, has also been proposed (Webb et al., 2007), but is seldom implemented in practice.

2.6.3 Rater Interpretation of Curriculum Objective and Test Item Content

Judges' alignment ratings "are highly dependent on a careful parsing of the content standards;" however, the "modal" state curriculum document may not have been developed "with sufficient care to support this level of parsing" (Beck, 2007, p. 130). Particularly if objectives are compound, partially duplicative, or insufficiently precise, rating is likely to be difficult (D'Agostino et al., 2008; Webb et al., 2007). Because objective statements are abstract, they may have "multiple legitimate interpretations," and potentially be translated into many different instructional practices (Hill, 2001, p. 302). Through interviews, surveys, and classroom observations of 25 Michigan teachers, Spillane (2004) found that even when teachers have similar familiarity with curriculum objectives, motivation to pursue the objectives during instruction, access to aligned curricular materials, and prior mathematics knowledge, they interpret a state's curriculum and test documents differently, and that these variations in interpretation influence their instructional decisions. During several alignment studies, Webb (1999) found that panelists, who included subject-matter experts and persons familiar with participating states' curricula and assessments, sometimes recognized that they were seriously uncertain about the intent of a particular objective, and were able to code it only after a state curriculum director provided guidance about its meaning (see also La Marca et al., 2000, p. 15). Even if a task statement has an unambiguous meaning, occasionally individual panelists may misinterpret it. Observing a curriculum development committee of teachers in one urban Northeastern school district, Hill (2001) reported that state curriculum objectives were sometimes misinterpreted by individual teachers. In some instances, even committee consensus decisions about district curriculum objectives partially reflected single teachers' misunderstandings when others either failed to offer a correction, or had the same

misunderstanding. Similarly, subject-matter experts have been observed to occasionally misunderstand the behaviors intended to be elicited by test items (D'Agostino et al., 2008).

Although careful selection of qualified subject-matter experts who have knowledge of the relevant curriculum documents should reduce the potential for rater misinterpretation of curriculum objectives or test items, classifications made by individual judges or the panel may be influenced by systematic bias. Alignment panelists may be too strict or lenient, tending to find too many or few matches, or to assign higher or lower ratings than warranted by tasks' content. While panelists should have some preexisting knowledge of the analyzed curriculum, or perhaps similar documents, "they should probably not have been heavily involved" in the development of either the curriculum or the test (La Marca, 2001, Methodological Considerations, para. 3), as such connections can positively bias their alignment judgments (Bhola et al., 2003). Sanford and Fabrizio (1999) observed that alignment panelists who had participated in test development exhibited "feelings of stress, frustration, and defensiveness" when their instruments were under review (p. 13). Curriculum alignment reviews conducted internally by test contractors may be particularly subject to bias. Buckendahl et al. (2000) concluded that employees of two test publishers found considerably larger proportions of test items aligned with Nebraska's English Language Arts curriculum goals than did review panels of classroom teachers, on average. Even if alignment panelists have not been involved in producing the documents under review, certain types of panelists may exhibit more lenient response sets. Bhola et al. (2003) cautioned that training for teachers participating in alignment needs to clearly define criteria for matching, in order to overcome their tendency to attempt to find objective matches for every item (Bhola et al.), or to match many items to multiple content topics (Herman et al., 2007). It has further been suggested that educator panelists who are, and are not, subject to a particular test-based

accountability system might tend to produce different judgments about alignment of tests to a particular curriculum (Roach et al., 2010), but the direction of any differences cannot be easily predicted because panelists' familiarity with the curriculum would also presumably vary.

Monitoring of judge comprehension during alignment review is limited by the goal of generating sets of independent ratings, and may vary across applications of the same alignment method unless there is a consistent procedure for allowing judges to seek clarification of document content. Hambleton (1980, pp. 211–212) recommended inserting known "bad" items, which do not measure any intended objective, into traditional content validation matching processes, rationalizing that the ratings of judges who matched a large proportion of decidedly off-topic items to particular objectives should be eliminated from any data analysis (see also Davis-Becker & Buckendahl, 2013), but recent published alignment methods do not include any such verification step. Adding a phase of discussion-based feedback regarding items for which there are serious discrepancies in initial content coding, analogous to the panelist group discussion sometimes facilitated during standard-setting procedures (e.g., Reckase & Chen, 2012), which does not force panelists to reach consensus on judgments about test items, could perhaps prevent gross misinterpretation of document content, as well as allow more monitoring of rater understanding by the moderator. The SEC alignment process includes group discussion of some items, but they are identified by panelists, rather than by the facilitator based on collected data (Porter et al., 2008). Alternatively, a feedback phase after initial coding could provide panelists with information about item writers' intent in constructing each item (Martineau et al., 2007). Current alignment methods assume rater competence, following training, to make the types of content classifications required, but tend to probe this assumption only through administration of exit surveys inquiring about judges' experience during the review

process (Davis-Becker & Buckendahl, 2013). Consistent, standardized analysis and reporting of participants' survey responses would permit users of alignment results to gauge the judges' understanding of the rating or matching task (Wyse & Viger, 2011), providing necessary evidence for validation (Davis-Becker & Buckendahl, 2013).

2.6.4 Rater Interpretation of Test Item and Curriculum Objective Cognitive Demand

Unlike item difficulty prediction, which requires raters to anticipate observable behavior (the response of an average examinee, or the center of the response distribution; e.g., Hambleton & Jirka, 2006), item cognitive complexity classification requires panelists to predict examinees' cognitive processing—the strategy they will tend to use to solve a problem—and then to judge the complexity of the processing requirements to execute that strategy. The training that occurs before alignment review guides judges to internalize the cognitive demand classification scheme utilized by a particular alignment method. To foster rigorous conceptualization of each item's response requirements, alignment procedures may instruct judges to complete each item, identifying the correct response prior to matching or rating its content (Ebel, 1956), and perhaps to assign corresponding objectives to each step of the solution process (Martineau et al., 2007). Training also usually includes cognitive demand coding practice using sample objectives or items (Webb, 2007).

Raters' understanding of the concept of item cognitive demand is shaped the content and delivery of specific instructions defining the concept, delineating its classification categories, and describing item features that should be considered in assessing cognitive demand. Wyse and Viger (2011) used a debriefing survey to probe item writers' understanding of cognitive demand following training on Webb's (1999) cognitive complexity rating scheme. The item writers included teachers and other educators, all of whom had at least three years of teaching

experience. The researchers interpreted some comments on the debriefing survey as evincing misconceptions about item cognitive demand. Particularly, many item writers seemed to conflate cognitive demand with item difficulty. However, most comments reflected understanding of at least some aspects cognitive demand that was consistent with the training provided.

After receiving training, and carefully considering task statements' features, judges may still find classifying tasks' cognitive demand to be challenging. In one Webb-type alignment review of an adult basic competency test and curriculum, Martone (2007) found that there was some disagreement among panelists' cognitive demand ratings for about two-thirds of the objectives, and for many of those objectives, initial ratings were nearly evenly split across two adjacent cognitive demand categories. Panelists' judgments about task cognitive demand are likely to be influenced by their understanding of the "developmental levels and prior instructional experience" of the examinee population (Herman et al., 2007, p. 121). For example, when reviewing high school mathematics test alignment, high school math teachers tend to rate the items' cognitive demand more highly than do university faculty (Herman et al.). Compounding the difficulty of predicting "average" cognitive demand, as items' cognitive complexity increases, students are more likely to use diverse processes (e.g., either algebraic or geometric reasoning) to reach the correct solution (Leighton & Gokiert, 2008), producing uncertainty about what objectives the items measure (Webb et al., 2007). Knowledge of the examinees' instructional experience may be particularly necessary to classify these items. Serious disagreement about many items' cognitive demand would suggest that more training is necessary to help panelists appreciate the meanings of, and distinctions among, cognitive demand categories (Martone, 2007).

2.7 The Relationship Between Test-Curriculum Alignment and Student Achievement Test Scores: Previous Empirical Findings

A basic premise of opportunity-to-learn research is that as students receive high-quality instruction following a particular curriculum, they learn, so their scores on test items (Schmidt, McKnight, Cogan, Jakwerth, & Houang, 1999; Wiley & Yoon, 1995) and tests (e.g., Schmidt et al., 2001) covering topics emphasized in the curriculum are expected to increase. Achievement test content validation arguments make the same claim, focusing on the role of the test: if test scores are valid measures of curricular attainment, truly reflecting the degree to which students have mastered the objectives, the scores should increase following relevant instruction (e.g., D'Agostino, Welsh, & Corson, 2007; Gulliksen, 1950). To guide the design of the International Association for the Evaluation of Educational Achievement's (IEA) cross-national mathematics studies, Travers and Westbury (1989) translated this theory into a model of curricular learning that distinguishes between the formal or informal curriculum intended by stakeholders in an educational system, the intended curriculum, and the instruction that students actually receive, the enacted curriculum. The intended curriculum is the content material that legislative authorities, such as national or state education agencies, intend for students to learn in school. The implemented, or enacted, curriculum is students' actual content exposure resulting from instruction during school. The attained curriculum is students' resulting content mastery, or achievement. Schmidt et al. (2001, p. 31) hypothesized that the intended curriculum might have not only an indirect effect on student achievement gains, mediated by instruction, but also a direct effect on gains. Because alignment indices are meant to reflect the degree of the correspondence between the intended curriculum, instruction, and the test instruments used to measure student achievement, under certain assumptions, the indices would be expected to be predictive of achievement gains.

If instructional quality is sufficient (e.g., La Marca et al., 2000), student motivation is adequate (e.g., McMaken & Porter, 2012), and test item scores (e.g., Muthén, Kao, & Burstein, 1991) or subtest scores (e.g., Schmidt et al., 2001) are sensitive to differences in instructional content, the strength of alignment between the test and curriculum, in conjunction with the amount of instructional time allocated to teaching the curriculum (Gamoran et al., 1997), should be positively related to student test score gains. Presuming these assumptions hold, the correlation between alignment indices and mean student test scores, or test score gains, could provide evidence of the indices' validity as measures of test-curriculum correspondence (Crocker et al., 1989) and of their potential utility for test developers and teachers (Webb, 2007). Considering the same structural relationships and assumptions from an OTL perspective, it would likewise be expected that if test-curriculum alignment indices are an indicator "of the potential of classroom instruction to influence student achievement" in a particular domain (Roach et al., 2008, p. 169), they should be related to student achievement gains (Schmidt & Maier, 2009). However, the strength of the relationship between alignment measures and achievement will likely be affected by the specific way that alignment is operationalized (Leinhardt & Seewald, 1981), as has been observed for OTL measures (Floden, 2002; Schmidt & Maier, 2009). Although the focus of this study is on alignment indices measuring correspondence between tests and curricula, to assess the extent that any alignment indices have been demonstrated to explain variability in student achievement or achievement gains, in the following sections we describe existing evidence for the impact of alignment between tests and curricula, instruction and tests, or instruction and curricula, on student achievement. Previous studies have variously represented student performance as total scores, subtest scores or item scores; all are reviewed here. We highlight results from mathematics, the subject area in which

most alignment-related research studies have been conducted, and which is most relevant to the present study, as well.

2.7.1 Instruction-Curriculum Alignment and Achievement Test Scores

Smithson and Collares (2007) studied the relationship between curriculum-instruction alignment indices and student achievement scores in underperforming Ohio schools (i.e., schools not making "adequate yearly progress" in students' average achievement, according to the state's ESEA criteria). They found that alignment indices were a statistically significant positive predictor of classroom mean achievement, controlling for grade level, although the effect size was less than one-quarter of a standard deviation in mean achievement scores. The effect remained after controlling for prior mean achievement, but the prior means represented scores from only about one-third of the students in the sample, so the coefficient was not expected to be an unbiased estimate of the population relationship between alignment and mean achievement gains. Using only the fraction of the student sample for which prior achievement scores were available, after controlling for economic disadvantage, grade level and prior achievement using a multilevel model, no significant relationship between teachers' instruction-curriculum alignment and students' achievement scores was observed.

In a random sample from the 10% of Ohio districts participating in the same instructional alignment study, Woolard (2007) found that elementary school buildings in ESEA "School Improvement" status generally reported lower mean teacher alignment scores in both math and language arts than buildings not in School Improvement status, although these differences were not statistically significant. He also found a small, significant positive correlation between schools' mean alignment and their annual mathematics Performance Index, a state-mandated

accountability indicator that was a weighted sum of each school's proportions of students in each proficiency category, by subject area.

Kurz et al. (2010) examined the relationship between instruction-curriculum alignment, calculated from SEC teacher questionnaire data, and classroom achievement averages of 18 volunteer general- or special-education Grade 8 mathematics teachers in an urban school district in Tennessee. Training was conducted according to established SEC protocols. Classroom-level correlations between the curriculum-instruction alignment index and mean achievement on Tennessee's summative state mathematics test were .64 for alignment of instruction reported at mid-year, and .58 average alignment reported at the end of the school year, relatively high. However, the authors cautioned that the correlation between alignment and achievement at the individual student level was likely to be considerably lower than the correlation at the classroom level. They recommended that future studies "should evaluate alignment alongside other known predictors of student achievement, including prior achievement, engagement, and other academic enablers" (Kurz et al., 2010, p. 142).

Polikoff and Porter (2012) studied the effects of instruction-curriculum alignment, as measured by the SEC index, on teacher "value-added" scores in 4th- and 8th- grade English language arts and math. The teachers surveyed were a self-selected subsample of teachers from the Measures of Effective Teaching study, which sampled teachers in six urban school districts. They were significantly more likely to be white, and had lower proportions of Black or American Indian students, than teachers who did not participate. Teacher value-added scores in a particular subject and grade level were calculated as average residuals from models of student achievement test scores that controlled for prior test scores and other individual student characteristics (several different achievement tests, including each student's state's achievement test, were

administered, and consecutively modeled as alternative outcome variables). Four measures of teachers' pedagogy based on student surveys or classroom observation protocols were also collected. The correlation between teachers' instruction-curriculum alignment index scores and the mean residualized achievement scores of their students was significant and positive in math, and larger than the correlations between any of the pedagogical measures and the value-added scores. However, after adding fixed effects for district-grade combinations and all the pedagogical measures as additional predictors of the value-added scores, the coefficient on math instruction-curriculum alignment became nonsignificant, although it remained positive. The authors interpreted their results, overall, to indicate that the SEC instruction-curriculum alignment index, or other content coverage measures derived from SEC data, might be predictive of teachers' average residualized student achievement scores, perhaps even more predictive than pedagogical measures, but suggested caution in interpreting the results due to several possible threats to replicability in the full study population, including possibly insufficient power, inadequate training of teachers prior to their completion of the SEC survey, or other irregularities in the subsample data.

2.7.2 Instruction-Test Alignment and Achievement Test Scores

Winfield (1993) surveyed 19 teachers of regular or supplemental 4th-grade mathematics regarding their relative instructional emphasis on the specific content of 68 sample items written to correspond to 12 mathematics objectives covered by an annual state achievement test. Because disadvantaged (i.e., Title I) students' scores on the achievement test were used by the school district to evaluate the effectiveness of schools' supplemental instruction for these students, the teachers would have experienced mild-to-moderate pressure to align their instruction to the test objectives. Teachers' responses to questions about "(1) the number of

times a mathematics concept was taught, (2) the frequency of review or re-teaching of the concept, (3) the number of settings in which the particular test format was used to teach the concept, (4) the frequency of usage of the format, (5) the extent to which the concept was emphasized in the school reading curriculum, and (6) the teachers' perception of student mastery of the concept" were used to produce a content emphasis scale score for each item (p. 292). Students in these teachers' classrooms who were eligible for Title I services then completed the test items. Analyzing the students' item scores, Winfield found that average content emphasis scale scores for each item for both the regular and supplemental teacher groups were moderately, positively and significantly correlated with item difficulty (i.e., *p*) values. That is, students were more likely to respond correctly to test items containing content that was emphasized during instruction.

A study by Gamoran, Porter, Smithson, and White (1997) is often cited as demonstrating that, in conjunction with instructional time, the alignment between instruction and a test instrument, as measured by the SEC index, predicts student achievement. Comparing achievement gains in three types of high school mathematics classes: general-track, transition, and college-preparatory, using a multilevel model the investigators found that "more rigorous content coverage accounts for much of the advantage of college-preparatory classes" over transition and general-track classes in math achievement gains (p. 325). The sample of 9th- and 10th-graders, drawn from four urban school districts in California or New York, was characterized as relatively low-achieving. For each participating classroom, the study calculated an indicator of content coverage that was a cellwise product of alignment, as computed from an SEC math content-cognitive demand matrix, and proportion of instructional time, as reported in a teacher survey. The model of achievement gains included covariates measured at both the

individual and classroom levels, but prior individual achievement was not among the predictors. Results indicated the indicator of content coverage was a marginally significant positive predictor of individual students' achievement gains over one school year. However, the authors cautioned that student achievement gains during the school year, which averaged 1.7 points on the 26-point test, may have been partially attributable to repeated administration of the same test form, and teachers reported expending, on average, only about 7% of instructional time during the year on content that appeared on the outcome test, raising questions about the test's suitability as an outcome measure. McMaken and Porter (2012) recommended that the Gamoran et al. (1997) study linking alignment to achievement gains should be replicated.

D'Agostino et al. (2007) investigated the impacts of 52 fifth-grade teachers' content emphasis, instruction-test alignment, and the interaction of these factors, on Arizona state mathematics achievement test scores. Teachers were asked to describe, in writing, how they taught two particular performance objectives from the state math curriculum, and provide sample classroom assessment items, if possible. Two subject-matter experts rated, on a three-point scale, the degree of alignment between teachers' instruction and items on the state achievement test that matched the two objectives. Teachers were also asked to report, on a four-point scale, the degree of emphasis they placed on each of 21 performance objectives, including 11 Grade 5 objectives. The correlation between teachers' emphasis and alignment scale scores was only .19, suggesting that these measures captured different aspects of teachers' practice. Controlling for individual student background variables including two math pretest scores, as well as for their schools' federal school meal program eligibility proportions, the authors used a multilevel model to predict fifth-graders' math achievement test scores from classroom level emphasis, alignment, and the emphasis-by-alignment interaction. Finding that both alignment scores and the

interaction between alignment and emphasis were significant predictors of math scores, the authors concluded that there was some evidence that students in classrooms where instruction was over-aligned to the test performed better than students in classrooms where instruction plausibly targeted curriculum objectives but not precisely as they were operationalized on the test. Students in both highly- and moderately-aligned classrooms performed better than those whose teachers described instruction that seemed inappropriate to foster achievement of the objectives. The authors cautioned that teachers' responses may have been influenced by desire to make their instruction appear aligned to curriculum, and that the true effect of instructional alignment on math achievement may have been confounded by positive relationships between alignment, and content and pedagogical knowledge, neither of which had been measured.

2.7.3 Test-Curriculum Alignment and Achievement Test Scores

Using different measures of test-curriculum alignment, or "overlap," studies in the 1980s yielded mixed results regarding the relationship between alignment and achievement test scores. The "curriculum" in these early studies was usually taken to be represented either by a textbook (Freeman et al., 1983), possibly with its ancillary instructional materials (Leinhardt & Seewald, 1981), or by a curriculum guide, a document outlining—or possibly detailing—content and performance goals for a particular course of instruction. To judge test-curriculum overlap, schools' degrees of curriculum-test match were rated by external curriculum experts (Mehrens & Phillips, 1986), or textbooks were systematically matched against a content taxonomy (Freeman et al., 1983; Mehrens & Phillips, 1987).

Mehrens and Phillips (1986, 1987; see also Phillips & Mehrens, 1988) conducted a series of studies to address the question of whether differences in schools' mathematics or reading curricula substantially affect student performance on commercial standardized achievement tests,

which were intended to assess elements common to school curricula nationwide. The authors (1986) used multivariate analysis of covariance to determine whether any variability in classroom mean subscores on an off-the-shelf standardized achievement test could be attributed to differential curriculum emphases across elementary schools in two Midwestern school districts. For reading and mathematics in Grades 3 and 6, district personnel used a 5-point scale to rate the degree of correspondence between the content emphases in each school's implemented curriculum, and in the test. The reading and math textbook series used by each school at the two grade levels were also recorded. After controlling for both mean pretest total scores and welfare eligibility rates in each school, neither test-curriculum correspondence rating nor textbook series used was a significant predictor of either mean total scores, or subscores, on the mathematics or reading tests among third- or sixth-graders. Although only 78 schools were included in the analysis, so statistical power was likely to have been low, even the adjusted mean test score differences among textbook series or test-curriculum correspondence rating categories were judged to be within the approximate classroom-level standard error of measurement for the scores. Using data from one of the districts, Phillips and Mehrens (1988) similarly found very small, nonsignificant differences in item p-values and objective-level (narrower) test subscores between curriculum-test content match rating groups, and textbook series groups, for both grade levels in both reading and mathematics. The authors cautioned that the district curriculum officers used as raters may not have been sufficiently knowledgeable regarding the curricula implemented in each school to judge test-curriculum correspondence.

Mehrens and Phillips (1987) used a 180-cell, three-dimensional matrix to classify the content of the Grade 5 and 6 math texts from three textbook series, which were used by different buildings in a school district, and the content of an off-the-shelf achievement test that was

administered annually in the district. Although the sequencing of topics differed across the textbook series, the cumulative content presented during Grades 5 and 6 was quite similar. The authors found that curricular emphasis proportion differences had no detectable relationship to item difficulty (*p*) differences computed from the scores of about 1,700 district sixth-graders who composed the three textbook groups. The average Rasch item difficulty value orders, and the mean item difficulty values for items covering similar content, also differed little for matched groups of students who used different textbook series. The authors concluded that the differences in curricula within a school district during that time period were not large enough to produce significant differences in standardized test scores.

2.8 Impact of Federal School Accountability Testing on Alignment

When most commonly-used alignment methods were developed, prior to the 2001 emendation of the ESEA, state curriculum documents varied widely in organization, level of specificity and grade level span (La Marca et al., 2000). Some curriculum documents were simple lists of content topics or of vague performance goals. Because these curriculum formats tended not to adequately specify cognitive demand, they hindered both the development of aligned tests and the alignment review process (La Marca, 2001). The amended ESEA required states to develop and disseminate written grade-level expectations, statements of relatively specific behavioral objectives for every grade level, reducing variation in curriculum document organization among states (Webb, 2007). While the previous ESEA emendation in 1994 had dictated that state accountability tests must match their curricula (Webb, 1997), few resources were devoted to ensuring, or even encouraging, compliance. Evaluation of proposed state accountability testing systems under the 2001 rendition of the law temporarily denied testing system approval to states that failed to submit alignment evidence (Schafer et al., 2009).

During the 1990s and early 2000s, "most states lacked a formal and systematic process" for determining the alignment between curriculum and assessments (Webb, 1997, p. 8). Some states expended little effort on alignment review; others recognized that their state achievement test corresponded poorly to the written curriculum, but lacked the resources to revise the curriculum or develop more appropriate tests (Wixson & Yochum, 2004). Alignment studies often deemed alignment between state-administered achievement tests and the relevant curriculum documents either to be low (Rothman, 2003), with item distributions concentrated on measuring the least cognitively-demanding objectives (Resnick et al., 2004; Webb, 1999), or to be inflated by the generality of many states' curriculum goal statements (Porter, 2002), each of which appeared to be measurable by a wide, content-diverse range of items. The tested curriculum (administered by states or school districts) was generally believed to have more influence than the written curriculum (developed by states) on the enacted curriculum (Glatthorn, 1999).

However, even in the decade before the amended ESEA took effect, activism by policymakers directed at controlling curriculum, instruction and assessment in some states appeared to influence teachers' instructional alignment, particularly in mathematics. Koretz (2008) describes the possibility of accountability-induced reallocation:

> Shifting of instructional resources (primarily instructional time, but other resources as well) among substantive parts of the curriculum to target better the particulars of the test. To some degree, reallocation is desirable, in that accountability tests are designed in part to signal what is important. Reallocation poses a risk, however, because tests are small and necessarily incomplete samples from the domains of achievement they are intended to represent. Allocating more time to one set of topics requires taking time away from

> others, and if the material that is dropped or de-emphasized is also important for the intended inferences about achievement, then scores can rise more than gains in achievement warrant . . . Numerous surveys have found that teachers report reallocating in response to testing. (p. 84)

In the spring of 2001, compared to teachers in states where student achievement tests had moderate or low stakes for teachers and schools, teachers in states where tests had high stakes reported being more likely to attempt to match the content and format of their classroom assessments to those of the state's achievement test (Pedulla et al., 2003). Efforts by states to shape classroom instruction may also have encouraged teachers to focus instruction on curriculum expectations. Controlling for an extensive set of state policy and school characteristics, as well as other features of eighth-grade math teachers' classrooms using a multilevel model, Swanson and Stevenson (2002) found that a state's level of "standards-based policymaking" (e.g., establishing curriculum objectives, often based on the National Council of Teachers of Mathematics' [NCTM] recommendations, administering curriculum-aligned assessments) was positively associated with the use of "standards-based instructional practices" (i.e., instructional content and practices recommended by the NCTM) in classrooms, with a "modest but substantively meaningful effect size" (p. 13).

The implementation of the amended ESEA, which increased the stakes of student achievement testing for schools in many states, has spurred public school educators to attempt to tailor classroom instruction to reflect state curriculum documents and assessment patterns. Recent research indicates that increased alignment with the curriculum is evident, particularly in elementary school mathematics instruction. Repeated annual surveys of educators from representative samples of California, Georgia, and Pennsylvania elementary and middle schools

by Stecher and colleagues (2008) between 2004 and 2006 documented changes to instruction attributed to the amended ESEA's accountability system. Most math teachers in all three states reported altering the content of their instruction to better reflect state curriculum objectives, although relatively few reported changing their proportional use of specific instructional strategies (e.g., direct instruction) over time. In spring of 2005, the middle year of the survey, about 75% of elementary math teachers, and a slightly smaller proportion of middle school math teachers, in the three states reported that they focused more instruction on tested topics than they would absent the high-stakes state test. Large percentages of elementary and middle school math teachers in all three states reported using item formats similar to those on the state test for classroom assessment more frequently than if the test had lower stakes. Many teachers also reported attempting to align their instruction to reflect the content of the state assessment; the lowest proportions of math teachers reporting such behavior were in California, where state policy prohibited public release of any test items from previous assessments. Results from the survey were similar in 2006.

As would be anticipated, efforts to increase instructional alignment to state curricula seem to have been concentrated in tested grades and subject areas. The amended ESEA mandated state achievement testing in reading and mathematics in elementary grades 3–8 beginning in 2005–2006. In Ohio, Woolard (2007) reported that school average curriculum-instruction alignment in mathematics, measured by the SEC index, rose markedly in Grades 2 and 3, the grades at or immediately before which accountability testing began, from a low base level in Grades K and 1. Science achievement testing is also required in two state-selected elementary grades, but it was not phased in until 2007–2008. Compared to science teachers in surveyed schools and school districts, math teachers have made more concerted efforts to align

the content of their instruction with state curriculum objectives (Stecher et al., 2008), and have achieved higher mean instructional alignment, as measured by the SEC index (Porter et al., 2007). However, there is little evidence of significant changes in instructional alignment among reading or English language arts teachers (Polikoff, 2012a).

Although many public school teachers reported attempting to increase alignment between the content of their instruction and state curriculum objectives, the magnitude of actual change in instructional content emphasis may have been small, and alignment may still be relatively low. Using alignment indices computed from SEC questionnaire responses collected from a selective sample of over 3,000 teachers from 23 states, Polikoff (2012a) concluded that Grade K–8 instruction-curriculum alignment increased slightly under the amended ESEA, with the most pronounced improvement in mathematics. Regression models of instruction-curriculum alignment change for the grade ranges K–2, 3–8 and 9–12 controlled for any time-invariant effects of particular states and grade levels on changes in instructional alignment. Over the six years between 2003 and 2009, the proportion of sampled math teachers' instruction that aligned to curriculum objectives increased by 3.8% for Grades K–2, and by 3.1% in Grades 3–8. Average instructional alignment over the study period and across the grades was low, however, with only about one-fourth of math instructional time distributed across content-cognitive demand combinations suggested by state curriculum documents. The sample did not depart wildly from national population average classroom characteristics, but was not claimed to be nationally representative, as most of the surveyed math teachers were from Indiana, Montana, Ohio, Oklahoma, or Oregon.

2.9 Summary of the Literature and Contribution of This Study

Alignment evidence is necessary for validation of state achievement test score interpretations. Although traditional methods of computing alignment exist, their application has usually compared test content to a test specifications table. Modern alignment procedures to compare test item and curriculum objective content differ along several dimensions, including the types of task features that are considered relevant to judging alignment, the reporting of error variability among panelists, whether an intermediate content classification table is used, and whether connecting items to objectives involves binary matching, rating, or both. There is little empirical support for the cognitive demand coding schemes adopted by modern alignment methods (Ferrara et al, 2011). Alignment methods' various indices and cutoff criteria espouse different definitions of test-curriculum alignment; none of the indices have cutoff values that have been devised based on empirical research, or are widely agreed upon (Davis-Becker & Buckendahl, 2013).

Alignment indices' validity as measures of test content representativeness depends on the conditions under which the rating data is collected. Monitoring of panelists' comprehension during alignment review tends to be limited by the desire to maintain the independence of their judgments. Although curriculum objectives may have multiple reasonable interpretations, panelists occasionally make clear errors of interpretation when decoding test items or curriculum objectives; however, these gross errors appear to be rare (D'Agostino et al., 2008; Hill, 2001). Panelists' judgments of task cognitive demand may be influenced by their understanding of the "developmental levels and prior instructional experience" of a given test-taker population (Herman et al., 2007, p. 121). As might be anticipated, panelists' findings may be biased if they have been involved in developing the tests or curricula under review. For this reason, while it is

usually recommended that judges have some previous familiarity with curricula that they will review, test or curriculum developers would not typically be recruited to alignment panels for state achievement tests. Evidence of variability across panelists' ratings of item and/or objective content indicates that ratings have been fairly or highly consistent during some alignment reviews, but that sometimes their consistency has been poor. Alignment index estimates may have acceptable reliability if reviews enlist 4 to 6 panelists; however, panelist numbers recommended by the Webb and SEC methods have been revised upward toward 5 to 8 to reflect the marked improvements in index reliability expected using data from additional panelists. The amount of rater disagreement averaged over to compute alignment indices in research or practical alignment studies has sometimes been high enough to warrant concern about the indices' meaning, but on most occasions when rater agreement has been reported, it has been acceptable. Overall, previous research provides some documentation supporting alignment indices' reliability and validity, but such evidence has not systematically been collected and reported.

Because alignment indices are meant to reflect the degree of content overlap between the intended curriculum, instruction, and the test instruments used to measure student achievement, under certain assumptions about instructional quality, student motivation and instructional sensitivity of the test items, the indices would be expected to be predictive of achievement gains. There is some evidence of a positive correlation between instruction-curriculum alignment and classroom or school mean achievement, particularly in mathematics. There is also some evidence that instruction-test alignment is a significant positive predictor of classroom mathematics achievement, and classroom and individual achievement gains. On the contrary, empirical evidence suggests that test-curriculum alignment is not significantly related to

classroom mean math achievement test scores. However, this conclusion reflects results from a single study conducted during the mid-1980s that analyzed an off-the-shelf, non-curriculum-based achievement test (Mehrens & Phillips, 1986, 1987). Further, the authors of the study cautioned that the district curriculum specialists engaged as raters may not have been sufficiently knowledgeable regarding curricula implemented in particular schools to accurately judge the extent of test-curriculum correspondence.

Modern alignment indices are an important warrant for claims in inferences (Kane, 2013) that generalize students' observed state achievement test scores to their expected performance on a universe of potential test tasks defined by their state's curriculum. To generalize curricular achievement test scores to performance under measurement conditions other than those observed, the task sample (i.e., item set) composing the test must be claimed to prompt behaviors representative of the activities listed in the relevant curriculum document. Such claims may be warranted by presentation of a particular alignment index or qualitative alignment evaluation as evidence of test content representativeness. The first purpose of this study is to seek external empirical backing for the alignment index warrants underlying some score interpretation validation arguments, as recommended by Davis-Becker and Buckendahl (2013). To investigate alignment indices' accuracy as measures of test content representativeness, I focus on checking two assumptions of the SEC alignment index formula: that counts of curriculum objectives are indicative of intended curricular emphasis (also an assumption of Webb's balance-of-representation alignment index), and that the cognitive demand categories adopted are best treated as nominal. The second purpose of this study is to probe the relationship between test-curriculum alignment and state average mathematics achievement.

CHAPTER 3: METHOD

My study uses data reflecting state math curricula from the 2005–2007 school years and student math performance in 2007. This time frame lies several years after passage of the amended ESEA, after the 2005–2006 deadline for states to fully implement its accountability provisions, and five years before any waivers of the accountability requirements were issued in 2012. Over the period from 2001–2007, considerable pressure on states led to increased uniformity in the organization, although not content, of state curriculum documents, and on teachers led to increased alignment between mathematics instruction and the written curriculum, providing a suitable context for testing test-curriculum alignment index function. Variability in curriculum topic-by-cognitive-demand coverage among and within states will contribute to the power of statistical tests of the overall relationship between content emphasis or alignment and test item performance, and results from different states should have at least some comparability due to increased similarity in the organization of curriculum documents. Student mathematics achievement and teacher instructional content emphasis data for this study is drawn from the National Assessment of Educational Progress (NAEP) 2007 and Third International Mathematics and Science Study (TIMSS) 2007, and measures of content emphasis for state mathematics curricula and the two achievement tests are taken from publicly-available SEC content analysis data.

Research Questions 1 and 2 propose examining the relationship between content emphasis proportions from SEC content matrices that represent state curriculum documents, and two types of external criteria: achievement test item performance and mean teacher-reported instructional content emphasis, across states, as validation evidence for the SEC alignment index. Both relationships are expected to be positive. In this study, I will use zero-order correlations

with instructional emphasis, and average marginal effects from regression models of item difficulty (item difficulty models, e.g., Gorin, 2006), as effect size measures to quantify the strength and direction of these relationships, if any. To estimate the unique effect of curricular content emphasis on test item performance, which is posited to also be influenced by many other item and examinee characteristics, the item difficulty models will control for item- and state-level characteristics believed to be among the most important. Research Question 3 asks if there is a statistically significant association between test-curriculum alignment, measured at the level of content topic, and mean test item performance (i.e., item difficulty) in a state. Alignment is expected to interact with curricular emphasis, such that its association with item difficulty becomes increasingly positive as emphasis on curricular content relevant to each particular test item increases. Research Question 3, like Research Question 1, will be investigated using an item difficulty model, although for a data at a different grade level. While the results of this study will, in any case, have to be interpreted with some caution due to the relatively small group of states with coded curriculum documents for the relevant time frame, they are expected to contribute evidence for validation of the SEC alignment index, and to quantify the relationship between state-level alignment and achievement during a time period when elementary mathematics teachers were under high pressure to target state curriculum objectives during instruction.

3.1 Data

It is reasonable to believe that the more different two compared curricula, "the more likely those differences will have an impact" on test scores (Mehrens & Philips, 1987, p. 358). State mathematics curricula show sufficient variation in objectives that it may be reasonable to expect differences in item-level achievement due to differences in opportunity to learn the

content. Reys et al. (2007) concluded that alignment of the curriculum objectives (i.e., "grade-level expectations") across the ten most populous US states was generally poor. Fourth-grade math objectives showed little consistency across the states examined—about one-quarter of grade-level expectations were unique to one state's curriculum document, while only about a third of objectives appeared in six or more states' standards. Similarly, a quantitative alignment analysis that coded the content emphasis of state curriculum documents using the SEC index (Porter et al., 2009) found relatively low alignment among states' K–8 math standards, particularly within grade, but also consolidating across grades. They determined that there was small common curriculum recommending instruction on particular number properties and basic operations in the early elementary grades (see also Reys et al., 2007), on estimation at most grade levels, on simple probability at Grade 7, and on providing interpretation of data displays at Grade 8.

I engage public-use SEC data on proportions of content coverage in state math curricula. I restrict my analysis to 11 SEC-participating states that neither adopted, nor made publicly available as drafts, any major curriculum document revisions during 2006, the year immediately prior to NAEP 2007 and TIMSS 2007 testing (with "NAEP" and "TIMSS" henceforth used to refer to the 2007 versions of these tests, unless otherwise noted, for brevity). Because content learned in previous grades is likely to impact performance, I will aggregate curriculum content emphasis matrices for the grade in which each test was administered with those of the previous grade (Mehrens & Phillips, 1987), yielding a matrix of proportions for each state. This unweighted summation assumes that "roughly the same 'amount'" of total curriculum content was covered in each grade (Porter et al., 2009, p. 264).

Both the NAEP and TIMSS studies assessed curricular mathematics achievement among fourth- and eighth-graders, collecting additional background information from sampled students, their math teachers, and school administrators. All US states participate in NAEP testing, and Massachusetts and Minnesota served as benchmarking participants for TIMSS. Using TIMSS and NAEP item responses, rather than state achievement test item responses, to measure academic mathematics achievement offers the advantages of cross-state comparability and the potential to control for factors, besides content coverage emphasis, hypothesized to affect test item performance. Although the content-cognitive demand categories implemented by the coarse-grained SEC matrix (MECG, 2004) and the two assessments' frameworks are not identical, the three schemes' content dimensions overlap heavily—all content categories appearing on NAEP and TIMSS were used during SEC coding—and the demand dimensions overlap partially. The SEC's content classification scheme will be mapped, separately, onto the two assessments' content coding categories.

3.1.1 SEC Data

Since 2001, researchers from the Wisconsin Center for Educational Research at the University of Wisconsin-Madison and the Surveys of Enacted Curriculum State Collaborative Project sponsored by the Council of Chief State School Officers have conducted or facilitated content analyses of curriculum documents and/or achievement tests from many states and school districts, as well as a number of national standardized tests (Porter et al., 2011). I engage public-use SEC data on proportions of content coverage in eleven states' math curricula, which would have been the active curriculum standards at, and prior to, the NAEP and TIMSS 2007 administrations. I restrict my analyses to states that participated in SEC alignment analyses, and neither adopted, nor made publicly available as drafts, any major revisions of their curriculum

documents during 2006: Alabama, California, Indiana, Kansas, Massachusetts, Michigan, Minnesota, New Jersey, Ohio, Oregon, and Vermont. Some of these states had relatively long-standing mathematics curriculum documents, while others' curriculum documents had been more recently introduced. Identification of states with stable curriculum documents during 2006 was based on consistent information from three nationwide policy reports that listed states' current curriculum documents, and dated and described any major published curriculum revisions occurring over the time intervals from 2005 to 2008 and 2005 to 2010 (American Federation of Teachers, 2008; Carmichael, Martino, Porter-Magee, & Wilson, 2010; Klein, 2005), with reference to state department of education websites for confirmation. NAEP 2007 and TIMSS 2007 test items have also been content analyzed using the SEC content classification scheme (Blank & Smithson, 2009).

The fine-grained SEC content matrix for mathematics, which is recommended for use in alignment analyses (Porter, 2002), has 915 cells. The coarser-grained version of the mathematics matrix, which consolidates specific content topics but retains the same cognitive demand distinctions as the fine-grained matrix, has 80 cells. Because achievement tests usually contain many fewer than 1,000 items, and test score users generally want to make inferences about performance on a domain broader than the "specific cells that happened to be tested" in a very large, detailed matrix, Mehrens and Phillips (1986, p. 186) cautioned that the coding matrix should not be too large. To allow classification of state curriculum documents' SEC content proportions according to the NAEP and TIMSS content categories, I will use proportions from the coarse-grained SEC state curriculum content analysis matrices, and NAEP and TIMSS assessment content analysis matrices for Research Question 3, as raw data. Because the accuracy of individual judges' ratings is likely to decrease as an alignment matching task requires more

detailed parsing of content items and knowledge of examinee behavior (Davis-Becker & Buckendahl, 2013), consolidating over detailed content topics may provide the best chance for a favorable assessment of alignment index validity. Sixteen content topics: Number Sense/Properties/Relationships, Operations, Basic Algebra, Advanced Algebra, Consumer Applications, Measurement, Geometric Concepts, Advanced Geometry, Data Display, Statistics, Probability, Analysis, Trigonometry, Special Topics (e.g., sets, logic), Functions, and Instructional Technology define the rows of the coarse-grained mathematics content matrix. Five cognitive demand types: Memorize, Perform Procedures, Demonstrate Understanding, Conjecture/Generalize/Prove, and Solve Non-routine Problems/Make Connections define the columns of the matrix. Descriptions of example response requirements that correspond to each cognitive demand type are listed in appendix Table A1. The proportion in each cell of the SEC matrix is taken to represent an estimate of the relative emphasis of that cell's content category by a test or curriculum document, based on panelists' item or objective content classifications (Porter, 2002).

Because knowledge is acquired cumulatively, and tests are often designed to measure knowledge that may have been taught in previous grades, investigations of the relationship between curriculum to test match and achievement should account for more than one grade's curriculum (Mehrens & Phillips, 1987). Kurz et al. (2010) interpreted findings from their small alignment study as suggesting that the relation between alignment and mean achievement becomes strong only when "students have been exposed to the instructional curriculum for a sustained period of time—in case of this study, for longer than 6 months" (p. 142). To account for the cumulative nature of knowledge acquisition in mathematics, I will aggregate the curriculum content emphasis matrices (e.g., Porter et al., 2009) for the grade in which each test

was administered (Grade 4 or Grade 8) with those of the previous grade (Grade 3 or Grade 7) by summing the matrix pairs and dividing each element by two. In one state, curriculum documents for the grade blocks 3–4 and 7–8, rather than for single grades, were coded; the proportions in these content emphasis matrices were taken to represent coverage in the relevant grade ranges.

3.1.2 Third International Mathematics and Science Study 2007 U.S. Benchmarking and National Assessment of Educational Progress 2007 Samples

The NAEP 2007 study assessed curricular mathematics achievement among fourth- and eighth-graders, collecting additional background information from sampled students, their math teachers, and school administrators (NCES, 2009). The survey used a two-stage stratified sampling design, selecting schools with probability proportional to size in the first stage, and about 30 students per sampled school in the second stage. All US states participated in NAEP testing between January and March 2007. In the states for which SEC curriculum content analyses are available, 37,689 fourth graders from 1,713 schools, and 35,182 eighth graders from 1,607 schools, in total, participated. NAEP sampled only public schools; Department of Defense and Bureau of Indian Education schools will be excluded from analysis. NAEP used a balanced incomplete block test booklet series design, so each student was administered only a fraction of the test item set. To fourth-grade students, 164 different items were administered; to eighth-grade students, 167 different items were administered. Test booklets were distributed in a spiraling manner so that the group of students receiving each item should, after accounting for unequal probabilities of selection, approximate a simple random sample from the population. In both grades, items covered five major content topics: Number Sense/Properties/Operations, Measurement, Geometry/Spatial Sense, Data Analysis/Statistics/Probability, and Algebra/Functions. Items were classified by the test developers as requiring one of three levels

of cognitive demand: Low Complexity, Moderate Complexity, and High Complexity, the definitions of which are provided in appendix Table A2. Counts of Grade 4 and 8 NAEP items in each content category are displayed in Table 1.

Like NAEP, TIMSS drew grade-based samples of students in their fourth and eighth years of formal schooling. Selected students were tested in mathematics and science. The US states of Massachusetts and Minnesota served as benchmarking participants, sampling large enough numbers of public school students to permit state-level achievement estimates to be obtained. These states' TIMSS sampling designs were based on the NAEP sample designs, and were specified to minimize duplicate selection of schools by the two studies at each grade level

TABLE 1
Distributions of NAEP 2007 Test Items by Content Category and Grade Level

Cognitive Complexity	Topic									
	Algebra		Data Analysis, Statistics, and Probability		Geometry		Measurement		Number Properties and Operations	
	Gr. 4	Gr. 8	Gr. 4	Gr. 8	Gr. 4	Gr. 8	Gr. 4	Gr. 8	Gr. 4	Gr. 8
Low	10	23	13	13	13	19	24	19	41	23
Moderate	9	19	6	13	9	12	11	9	21	14
High	1	3	1	0	1	0	0	0	2	0

Source. National Assessment of Educational Progress 2007.

TABLE 2
Distribution of TIMSS 2007 Grade 4 Test Items by Content Category

Cognitive Domain	Topic		
	Data Display	Geometric Shapes and Measures	Numbers
Applying	11	25	32
Knowing	6	23	39
Reasoning	9	9	20

Source. Trends in International Mathematics and Science Study 2007.

TABLE 3
Distribution of TIMSS 2007 Grade 8 Test Items by Content Category

Cognitive Domain	Topic				
	Algebra	Chance	Data	Geometry	Numbers
Applying	15	5	13	27	25
Knowing	32	5	9	8	27
Reasoning	17	0	8	12	8

Source. Trends in International Mathematics and Science Study 2007.

(see Olson, Martin, & Mullis, 2008, for a detailed description of the sampling procedure). Testing took place between March and June 2007. In total, 3,593 fourth graders representing 97 schools, and 3,674 eighth graders from 97 schools, participated in these two states. Fourth-graders were administered 179 different mathematics items, while eighth-graders completed 215 different items. TIMSS also used a balanced incomplete block test booklet series design. At Grade 4, items covered three major content topics: Numbers, Geometric Shapes and Measures, and Data Display. At Grade 8, items covered four major content topics: Numbers, Algebra, Geometry, and Data and Chance, which included statistics and probability items, the "Chance" subtopic. Items were classified by the test developers as requiring one of three types of cognitive demand: Know, Apply, and Reason, the definitions of which are provided in appendix Table A3. Counts of Grade 4 and 8 TIMSS items in each content category are displayed in Tables 2 and 3, respectively.

3.1.3 Comparison of TIMSS and NAEP Assessment Frameworks, and SEC Content Coding Categories

Although the proportions of items covering each content type and specific objectives to be assessed differ between NAEP and TIMSS at each tested grade level, the tests' items at each grade level cover an identical set of broad content areas (Neidorf, Binkley, Gattis, & Nohara, 2006). The mathematics achievement conceptualizations and target task domains of NAEP and

TIMSS have wider scope or referent generality than most state math achievement tests. However, NAEP, TIMSS, and particular states' achievement tests would be expected to have some item types in common. Further, research suggests that mathematics curriculum interventions, if implemented with reasonable fidelity, can produce sizable score gains on subject-area achievement tests, even if the tests have not been intentionally aligned with the new curriculum units (e.g., Senk & Thompson, 2003). Although the targets of inference in the TIMSS and NAEP studies are broad constructs, I will treat their item sets as corresponding to a potential state curriculum that defines an observable mathematics achievement trait, and interpret students' item performance as representing "relative degree of content acquisition" (Haertel, 1985, p. 24).

The content and cognitive demand categories of the SEC content language and the mathematics curriculum frameworks used to classify nations' curriculum materials in the early TIMSS studies (e.g., Robitaille et al., 1993) are similar. Both document analysis systems use a common classification scheme to describe the academic content of curricula and tests (Webb, 1997), and describe cognitive demand categories representing distinct types of observable behaviors (e.g., communicating, solving routine procedures) that are not assumed to be ordered and do not imply any particular underlying assumptions about examinee cognitive processes. More recent TIMSS assessment frameworks, including that used for TIMSS 2007, have been developed from the TIMSS 1995 curriculum framework (Mullis et al., 2005). Compared to the TIMSS 1995 curriculum framework, the TIMSS 2007 assessment framework uses fewer, more general content categories, and cognitive demand, rather than observable performance, categories (Mullis et al., 2005), more similar to the NAEP 2007, but less similar to the SEC, coding scheme than was the TIMSS 1995 framework.

As appropriate based on typical elementary school mathematics curricula and the frameworks of the NAEP and TIMSS assessments, the two tests cover only a subset of the SEC content topics. The SEC content topics of Advanced Algebra, Consumer Applications, Advanced Geometry, Statistics, Probability, Analysis, Trigonometry, Special Topics, and Instructional Technology are not covered on TIMSS at Grade 4. The SEC content topics of Consumer Applications, Analysis, Trigonometry, Special Topics, and Instructional Technology are not covered on TIMSS at Grade 8. The SEC content topics of Consumer Applications, Analysis, and Trigonometry are not covered on NAEP at Grade 4. The SEC content topics of Consumer Applications, Analysis, and Trigonometry are not covered on NAEP at Grade 8. The cognitive demand dimension of the SEC content matrix appears to have some overlap with the cognitive demand categories of NAEP and TIMSS. Specifically, all three classification schemes group together items that require extended reasoning to solve non-routine problems. NAEP explicitly describes its High Mathematical Complexity items as intended to require more demanding cognitive processing than items in other categories (NAGB, 2006), but TIMSS and the SEC do not make this claim about items in their Reasoning, or Conjecture/Generalize/Prove or Solve Non-routine Problems/Make Connections, categories, respectively.

3.2 Models

If test scores are measuring the intended trait, differences among the anticipated response processes used for, and content of, the test's items should explain some of variability in average item responses, proportion-correct item difficulty (Gorin, 2006). Item difficulty models specify particular item characteristics that are believed to affect examinees' average probability of correct response. Traditional item difficulty analysis models (e.g., Bejar, 1993) regress hypothesized important task features on the classical difficulty parameter for each item. To

determine the extent to which the knowledge and skills that affect observed item difficulty match difficulty features intended by the test developer, the proportion of variability in item difficulty explained by the modeled item features (i.e., an R^2 value), and effect sizes for each factor, are usually examined (Gorin, 2006). Particular item features "typically show similar relationships" with classical proportion-correct and item response model item difficulty parameters (Mislevy, Steinberg, & Almond, 2002, p. 122), which are less-commonly modeled (Gorin, 2006). Other approaches to item difficulty modeling analyze individual examinee response data using specialized Rasch item response models (e.g., Fischer, 1997), or latent class models (e.g., Tatsuoka, Corter, & Tatsuoka, 2004). Embretson and Daniel (2008) noted that coefficients estimated from individual data for each hypothesized difficulty factor would be consistent and expected to be unbiased, and would tend to have smaller standard errors than coefficients in analogous models estimated from estimated item parameter values. However, in an empirical study, they found that the magnitude and direction of coefficients for each item feature associated with difficulty were similar, regardless of whether they were estimated from individual data or Rasch item difficulty statistics.

Among item surface features or response process characteristics posited to affect mathematics item difficulty, early studies of Graduate Record Examination mathematical reasoning items suggested that cognitive complexity ratings were the most consistently useful predictor of classical (Chalifour & Powers, 1989) or item response model (Enright & Sheehan, 2002) item difficulty, and that structural features of items including the number of assignments to position that were fixed for the elements to be manipulated in the problem, and the amount of information from the rules and conditions that was actually required by the intended solution process were also significantly related to item difficulty (Chalifour & Powers; Enright &

Sheehan), as was one linguistic feature: verbal load—the number of words in the prompt (Chalifour & Powers). Recent studies of math achievement test items from state testing programs indicate that linguistic features may be an additional important determinant of item difficulty for elementary school students, as Abedi and Lord (2001) contended. Shaftel, Belton-Kocher, Glasnapp, and Poggio (2006) concluded that elementary students' probabilities of responding correctly to math test items appear to be primarily influenced by structural, trait-relevant problem features, particularly if the test development process has been rigorous. Evidence clearly suggests that classical item difficulty (i.e., item easiness) decreases with increased inclusion of mathematics vocabulary terms (Ferrara, Svetina, Skucha, & Davidson, 2011; Shaftel et al., 2006), a linguistic but trait-relevant item feature. However, certain purely linguistic features, particularly the number of ambiguous words in the item stem or response options, also significantly impede item performance (Ferrara et al., 2011; Shaftel et al., 2006), and there is no evidence that the Webb depth of knowledge or NAEP mathematical complexity cognitive demand coding schemes are predictive of item difficulty in the middle elementary grades (Ferrara et al., 2011). Noting that task models containing only item features seldom explain much of the variability in item difficulty values, Ferrara et al. (2011, p. 13) hypothesized that some of the additional variation in item difficulty is attributable to differences in opportunity to learn the item content, and suggested the future item difficulty studies should model OTL.

Ability estimates of students' standing on the overall mathematics achievement trait measured by each test, at the time of testing, would be anticipated to capture much of the effect of previous, aligned instruction on students' item performance, suggesting that latent variable modeling using student-level response data would not be ideal to answer my research questions, and that modeling of observed item difficulty values might be preferred. In this study, I will take

classical item difficulty (i.e., proportion-correct, "p") values for each state-item combination as the outcome variable. Examining the item difficulty distributions for evidence of severe non-normality using histograms, and skewness and kurtosis measures, and a D'Agostino-Pearson K^2 test (D'Agostino, Belanger, & D'Agostino, 1990), I found that all four difficulty value distributions were appreciably non-normal. Those for the Grade 4 NAEP and TIMSS and Grade 8 TIMSS items were slightly negatively skewed and had low kurtosis, and could be rendered approximately normally distributed by a logit transformation (e.g., Cox & Snell, 1989). The Grade 8 NAEP item difficulty values, however, followed a somewhat heavy-tailed distribution that could not be normalized by any power transformation (results of a Box-Cox computation indicated that the optimal transformation exponent to normalize the distribution as nearly as possible was 1.03—essentially, no transformation).

Rather than utilize ordinary least-squares estimation and a linear regression model for the raw or logit-transformed item difficulty values, I will use maximum likelihood estimation to estimate fractional logit regression models at each grade level. The so-named "fractional logit" model (Papke & Wooldridge, 1996) is a generalized linear model with a logit link function and Bernoulli variance function that is often used in econometrics applications when the dependent variable is a proportion. Compared to traditional item difficulty models, fractional logit models are advantageous in that they do not require the distributional assumptions of ordinary least-squares regression (e.g., continuity, normality of the population error distribution) that are unlikely to be met by item difficulty values, they permit observed item difficulty values anywhere on the closed interval between 0 and 1 (including from items that all students within a state answer correctly or incorrectly, which occur occasionally in real item data) and also produce predicted values in the unit interval, and they can yield an interpretable effect size

measure (Wooldridge, 2010) under assumptions that are generally more likely to be plausible than those of ordinary least squares. Since the heteroskedasticity-robust sandwich estimator for the standard error of the fractional logit model regression coefficients is consistent even when the Bernoulli variance assumption fails, as recommended by Papke and Wooldridge (1996), sandwich standard error estimates will be used for inference from these models, which will be implemented using the software Stata.

Although mathematics test data from two grade levels is available, to reduce the uncertainty interpreting in statistical test results that would be caused by multiple testing, I will utilize the Grade 4 data from NAEP and TIMSS to investigate Research Questions 1 and 2, and the Grade 8 data from both assessments to pursue Research Question 3. For individual state item difficulty models, predictors will include the relevant SEC curriculum content emphasis proportion, mean teacher-reported instructional emphasis on the item's topic, and the item's cognitive category. In overall cross-state models, I will include additional measures of state characteristics that are posited to affect mean item performance and potentially correlated with curricular emphasis proportions. Further, in the Grade 8 cross-state models, I will also add the pertinent state's mean Grade 4 2003 NAEP scale subscore on each item's content topic to control for prior achievement in the tested cohort. As additional external validation evidence, I will examine the anticipated positive relationship between alignment index curricular content emphasis proportions and a more proximal measure, mean teacher-reported content emphasis, estimating the correlation between content emphasis proportions and mean teacher emphasis by topic across states.

3.2.1 Models for Research Question 1

To address my first research question, regarding whether unweighted counts of curriculum objectives can be considered indicative of intended content emphasis in a particular curriculum document, I will examine validation evidence from concurrent measures: partial regression coefficients representing the unique relationship between transformed counts—proportions—in each cell of a curriculum content matrix and other variables that measure curricular content emphasis, or are expected to be positively correlated with content emphasis.

The NAEP Grade 4 data contain a measure of instructional content emphasis: mathematics teachers' self-reported ratings of their emphasis of each major content topic tested during instruction of the sampled students, making it possible to determine if there is any relationship between topic emphasis proportions from the SEC and mean reported instructional coverage of that content. For this analysis, because teachers were asked about topic emphasis, I will collapse over cognitive demand categories of the SEC matrix to generate a total emphasis proportion for each content topic. Content topic proportions will then be aggregated as necessary to correspond to the major content topics used for reporting by NAEP. Since NAEP samples students, not teachers, mean teacher ratings for each state will be computed as means of emphasis in instruction received by individual students, accounting for the sampling weights. Using NAEP Grade 4 teacher survey data for all nine states, the Pearson correlation between teachers' mean instructional emphasis of NAEP Topic i (i = 1, 2, . . ., 5) in State k, and the corresponding residualized SEC content emphasis proportion for that state will be computed. Because students' teachers are not randomly assigned to states, prior to computing the correlation, variability in curriculum content emphasis proportions attributable to state characteristics will be removed from the proportion measure by regressing it on four principal

components scores on a set of State *k* educational characteristics, which are described in more detail subsequently. The computed correlation between curricular and instructional emphasis will be equivalent to the standardized coefficient α from the regression model depicted in Equation 1:

$$E(Y_{ik}|X_{ik}, \mathbf{Z}_k) = \alpha X_{ik} + \boldsymbol{\gamma}\mathbf{Z}_k \,, \qquad (1)$$

where Y_{ik} is mean instructional emphasis on Topic *i* in State *k*, X_{ik} is the SEC content emphasis proportion corresponding to Topic *i* row in State *k*, and $\mathbf{Z}_k$ is a matrix of four principal components scores on a set of State *k* educational characteristics, described in detail following presentation of Equation 2 below.

If SEC curriculum content matrices are a reasonable representation of the content topic emphases in the intended curriculum, and instruction follows the curriculum, cellwise emphasis proportions should be positively correlated with mean teacher content emphasis survey responses. This analysis will also contribute to checking one of the assumptions of the third research question—that instruction largely follows the curriculum. Although the TIMSS Grade 4 data also contains a measure of instructional content emphasis: math teachers' reports of the proportion of instructional time devoted to each of the three major TIMSS content topics, a correlation estimated from mean instructional emphasis by topic in only two states was unlikely to be stable or have any generalizability to the US population, so it was not computed, but the mean instructional content emphasis measure was used as a covariate in additional analyses described subsequently.

After examining correlations between teachers' mean content topic emphasis and curriculum content topic emphasis, I will examine the relationship between item difficulty and curriculum content emphasis, represented by the proportion of objectives in the corresponding

SEC matrix cell. Because item performance is hypothesized to be affected by instruction at previous grade levels, as well as at students' current grade level, for each state, the SEC curriculum content matrix for each tested grade level will be aggregated with the matrix for the previous grade level. Again, these aggregated curriculum content matrices will be collapsed across some topics so that content emphasis proportions correspond to the content topic categories used by NAEP or TIMSS, as appropriate. For this analysis, to retain the assumption of the SEC alignment index that cognitive demand categories are nominal, representing different types, but not levels, of required cognitive processing for tasks (which admittedly is possible only to a limited extent), while permitting comparability to the NAEP and TIMSS cognitive demand categories, I consolidate some SEC content emphasis proportions within topic, combining the cognitive demand categories Memorize, Perform Procedures, and Demonstrate Understanding, and likewise the categories Conjecture/Generalize/Prove, and Solve Non-routine Problems/Make Connections, but maintain the distinction between curriculum objectives that require extended reasoning, and those that do not.

To estimate each item's classical difficulty parameter, the mean response on each binary item, I will conduct a subpopulation analysis of the item response data by state that accounts for each assessment's complex sampling design features. Because not every test item is administered to each student, taking the estimated average probability of correct response to each item from the sample, with cases weighted by the sampling weights, as an estimate of statewide probability of correct response relies on systematic random sampling of students within selected classrooms to take each item, produced by spiraling distribution of the different test booklets within each classroom. However, some students who were randomly administered a particular test item may have failed to respond, or produced an unscoreable response. Omitted items will

be scored as wrong; not reached items will be assumed to be missing completely at random (i.e., MCAR). All items that had identified correct answer(s) were included in my sample, with the exception of a small number of NAEP items that were scored as clusters due to reported high error correlations. From these item groups (2 clusters of 2–3 items each in both Grades 4 and 8), because the fractional logit model assumes item difficulty observations are independently distributed, which is not true for these items, and because the content topic and cognitive complexity were not recorded for the overall cluster response, the first item in each cluster was included in the sample. Both NAEP and TIMSS include some open-ended items with maximum scores greater than 1. In addition, for some of the NAEP open-ended items, scores for up to three raters are reported. Throughout my analyses, the proportion of students who earn the maximum score for a fully correct response from the majority of the raters (when applicable) will be treated as the item difficulty.

To model variability in the probability of correct item response perhaps attributable to differences in curriculum exposure across the examinee population, I will use fractional logit models with the conditional mean of states' estimated classical item difficulty values as the outcome, as shown in Equation 2,

$$E\left(Y_{jk}\middle|X_{jk},\mathbf{W}_j,\mathbf{Z}_k\right)=\Lambda(\alpha X_{jk}+\boldsymbol{\beta}\mathbf{W}_j+\boldsymbol{\gamma}\mathbf{Z}_k), \qquad (2)$$

where Y_{jk} is the estimated Grade 4 item difficulty (*p;* proportion fully-correct responses) for Item *j* in State *k*, X_{jk} is the SEC content emphasis proportion corresponding to topic-by-cognitive demand cell of Item *j* in State *k*, $\mathbf{W}_j$ is a matrix of Item *j* characteristics (NAEP or TIMSS classification), $\mathbf{Z}_k$ is a matrix of State *k* educational characteristics, which include mean teacher-reported content emphasis measures and a set of four principal components scores, and $\Lambda(\cdot)$ is the logistic function. In models for the TIMSS data, which as available for only two states, $\mathbf{Z}_k$

will be replaced by a single state indicator dummy. Because the instructional content emphasis measure is hypothesized to be a potential mediator between curricular content emphasis and achievement outcomes (e.g., Travers & Westbury, 1989), it will be introduced into the model last; results both including and excluding this predictor will be reported.

Most NAEP and TIMSS items are not publicly released, so item characteristics available as task model variables are limited to those found in published information, which include cognitive demand ratings and content topic codes generated by the test developers, but not linguistic feature codes. To account for state-specific differences in item performance attributable to variation in state educational characteristics that were potentially correlated with states' decisions about curricular content emphasis, means of some of these variables for each of the 50 US state populations were either obtained from the Digest of Education Statistics (NCES, 2008) or computed from the Grade 4 (or Grade 8) NAEP student background data. State means drawn from the Digest included the percentages of adults holding bachelor's or advanced degrees in 2006, of children living in poverty in 2007, and of children suspended or expelled from public schools during school year 2005–2006, as well as median income in 2005 and per-pupil expenditures in school year 2005–2006. Means estimated from the NAEP background data, and specific to the Grade 4 (or Grade 8) student population included the percentages of minority students, of English Language Learners, of students attending schools in rural areas or small towns, of students who were above the average age for their grade, of students eligible for the federal school meal program, of students who had transferred to their present school within the last school year, and of students who had computers at home, as well as students' mean number of absences in the past month and score on a scale measuring the frequency with which students talked to their parents about schoolwork. State mean NAEP 2007 grade-level reading

scale scores (NCES, 2013) were also tabulated. Unfortunately, some variables possibly related to both state curricular emphasis and students' test item performance, particularly mean mathematics instructional time, were measured by the NAEP teacher questionnaires but were missing for 15% or more students at both grade levels, with similar proportions of missing responses across states. This level of missingness, which was unlikely to be completely at random, was deemed too high to yield accurate estimates of state mean math instructional time.

Because this collection of state-level variables was not of main interest in my analysis, to reduce the number of parameters that had to be estimated in the instructional emphasis and item difficulty models, principal component analysis was used to determine the linear combinations of variables in this set that would capture most of their variability. Since the state educational variables were measured on many different scales, prior to conducting PCA, all variables were standardized. Because one state, Alaska, had suppressed student responses to most items in the NAEP background questionnaires, yielding an incomplete raw data matrix of state variable means, rather than conducting PCA of the raw data matrix, I analyzed the EM-estimated covariance matrix (in this case, a correlation matrix) that could otherwise provide starting values for imputation using a multivariate normal regression model, as suggested by Truxillo (2005). The first four eigenvalues of the estimated correlation matrix for the state fourth- (and also eighth-) grade population means were greater than one (e.g., Jolliffe, 2002), the average value of the matrix's eigenvalues. (The first four eigenvalues of both observed correlation matrices, omitting Alaska from the dataset, were also greater than one, and principal components scores calculated from weights determined by PCA of these matrices were each correlated in excess of .99 with their corresponding scores obtained from analysis of the EM-estimated correlation

matrix in the 49 states with complete data.) The first four principal components explained more than 75% of the variance in the original variable sets for both grade levels.

For both grade levels, PCA results showed that the first two eigenvalues of the correlation matrices were greater than two, and the first two principal components had interpretable patterns of weights, while the third and fourth principal components had weights that would result in somewhat less meaningful scores. The first principal component had similar patterns of weights at both grade levels. Both Component 1s had weights with absolute values greater than .3 for five variables: the percentage of children living in poverty, of adults holding at least a bachelor's degree, of students eligible for the federal school meal program, and of students with a computer at home, and the mean NAEP 2007 reading score; the Component 1 scores could be interpreted as measuring the mean household SES of students in each state; the scores *decrease* with increasing mean SES. In the fourth grade data, Component 2 had weights greater than .3 for four variables: the percentage of minority students, of English Language Learners, of students attending schools in rural areas or small towns, and of over-age students; the Component 2 scores could be interpreted as increasing with the heterogeneity of a state's student population. In fourth grade, Component 3 loaded most highly on four variables: the percentage of transfer students, of children living in poverty, and of children suspended or expelled from school, and the mean number of absences from school; Component 3 scores may capture elements of a state's school disciplinary climate; they increase with the percentage of suspensions and expulsions. In fourth grade, Component 4 loaded most heavily on three variables: expenditures per pupil, the percentage of over-age students, and the mean number of school absences; Component 4 scores primarily measure per pupil expenditures, and increases with expenditures, but also with mean absences.

In the eighth grade data, the same variables weighted highly on Component 2 as in the fourth grade data, but their signs were in opposite directions, so the Component 2 scores should be interpreted as increasing with the homogeneity of a state's eighth-grade student population. Component 3 loaded most heavily on the mean number of absences, the percentage of Grade 8 transfer students, and the mean frequency of discussing school at home; Component 3 scores appeared to primarily measure student mobility, and increased with student mobility. In eighth grade, Component 4 again loaded most heavily on per pupil expenditures, followed by the percentages of English language learners, transfer students, and students suspended or expelled from school; Component 4 scores gave per-pupil expenditures the largest positive weight, but percentage of suspended/expelled students also received a sizable positive weight. To retain capacity to account, in the research models, for differences in state educational climates that could affect both curriculum development and students' academic outcomes, state scores on all of the first four principal components were computed, and those from the 11 states with SEC curriculum content emphasis data were saved for use as control variables in the item difficulty models.

Because the magnitude of the relationship between curricular emphasis measures and item difficulty may vary substantially across content topics or across states, given the exploratory nature of this study, after obtaining the cross-state results, I will estimate the model within states, and within content topics, dropping the state-specific principal components scores or content topic indicators, respectively, from the model.

3.2.2 Models for Research Question 2

To address my second research question, I will consider the change in the estimated relationship between the SEC content emphasis proportion variable and Grade 4 NAEP or

TIMSS item performance when the emphasis proportions account for content at or above the cognitive demand level of a particular item, rather than only content at the cognitive demand level of the item. That is, items in the consolidated cognitive demand category that is hypothesized to represent a lower level of demand will be assigned the total curriculum content emphasis proportion for their topic, including the quantity from coverage of hypothesized more-demanding curriculum objectives in the Conjecture/Generalize/Prove, and Solve Non-routine Problems/Make Connections categories. Then, simple and multiple regression coefficients for the SEC content emphasis proportion variable as a predictor of state-item difficulty values will be re-estimated by topic, and overall. Because these estimates are not independent of (and in fact are highly dependent on) those obtained previously, it will not be possible to perform formal statistical tests of the differences, and so the differences will be inspected and described qualitatively. If cognitive demand categories are ordered, instruction follows the curriculum, and instruction on more demanding content related to the same topic benefits performance on items with less demanding content, I expect that the correlations should increase in the positive direction.

3.2.3 Models for Research Question 3

My third research question asks if differences in test-curriculum alignment can explain any of the cross-state variability in students' item performance on NAEP or TIMSS at Grade 8. Gamoran et al. (1997) suggested that the effect of instruction-test alignment (the "configuration of coverage") on achievement gains should depend on the amount of instructional time devoted to tested topics (the "level of coverage;" pp. 330–331). They recommended using the product of content emphasis and alignment, operationalized using an SEC-type measure, to predict achievement gains. D'Agostino et al. (2007) modeled instruction-test alignment relationships

with students' state math test scores using indicators of instruction-test alignment, content emphasis, and the product of these two variables. I will combine these two approaches. I will use a measure of cellwise topic-by-cognitive demand alignment based on SEC content analyses of state curriculum documents and the NAEP and TIMSS 2007 tests, given in Equation 3,

$$1 - \left|\pi_{X_{i,j}} - \pi_{Y_{i,j}}\right|, \quad (3)$$

where $\pi_{X_{i,j}}$ denotes a cell proportion in a state curriculum content matrix *X* and $\pi_{Y_{i,j}}$denotes the corresponding cell proportion in the NAEP or TIMSS content matrix *Y*. This measure will be bounded between 0 and 1, inclusive, with higher values intended to indicate better alignment between the coded assessment and curriculum document for a particular content cell. I will also include as predictors the emphasis proportion for that content cell from the curriculum document, and the interaction between alignment and curriculum content emphasis.

An alternative measure of alignment is potentially available in the TIMSS data. TIMSS conducted a "test-curriculum matching analysis" during which one or more persons familiar with each particular jurisdiction's intended curriculum determined whether each TIMSS math test item was covered in the curriculum at that grade level (for more than half of the students in the jurisdiction, if the intended curriculum varied), or not (Mullis, Martin, & Foy, 2008, p. 439). The judge(s) from Massachusetts determined that all the math items in the TIMSS Grade 8 test were intended to be taught at that grade level, while the judge(s) from Minnesota found that the content of 2 items would not have been covered at that grade level by instruction that followed their state curriculum. Due to the very limited variability on this binary matching measure, I will not endeavor to compare it to the proportions underlying the SEC alignment measure.

To model the conditional mean of Grade 8 NAEP or TIMSS state item difficulty, predictors will include alignment and content emphasis indicators from the SEC data, mean

teacher-reported instructional emphasis on the item's topic, and a set of item characteristics indicators, as shown in Equation 4,

$$E\left(Y_{jk}\middle|X_{jk}, U_{jk}, \mathbf{W}_j, \mathbf{Z}_k\right) = \Lambda(\alpha X_{jk} + \tau U_{jk} + \theta X_{jk}U_{jk} + \boldsymbol{\beta}\mathbf{W}_j + \boldsymbol{\gamma}\boldsymbol{Z}_k), \qquad (4)$$

where Y_{jk} is the estimated Grade 8 item difficulty (*p;* proportion fully-correct responses) for Item j in State k, X_{jk} is the SEC content emphasis proportion corresponding to topic-by-cognitive demand cell of Item j in State k, U_{jk} is the SEC alignment (to NAEP or TIMSS) measure corresponding to topic-by- cognitive demand cell of Item j in State k, $\mathbf{W}_j$ is a matrix of Item j item characteristics (NAEP or TIMSS classification), $\mathbf{Z}_k$ is a matrix of State k educational characteristics, which include the NAEP 2003 Grade 4 scale subscore corresponding to the topic of Item j in State k and a set of four principal components scores, and $\Lambda(\cdot)$ is again the logistic function. Neither NAEP nor TIMSS collected information from Grade 8 teachers regarding instructional emphasis by topic, so this potential mediator between curricular emphasis and item difficulty cannot be modeled. Analogously to the state educational characteristics data for the Grade 4 population, the matrix of State k mean educational characteristics for the Grade 8 population is reduced to a matrix of principal components scores, as detailed previously.

If the written curriculum has little overlap across years (Smithson & Collares, 2007), and instruction corresponds well to the curriculum, so that the product of test-curriculum alignment and content emphasis measures new, recent opportunities to learn the test content, this interaction should be more predictive of achievement gains than of cross-sectional achievement scores. In the cross-state models, I use each state's mean Grade 4 2003 NAEP score as a pretest math achievement measure for the tested cohort. Because the importance of the alignment-emphasis interaction for explaining variation in test item difficulty may vary across content topics or alignment data collection events, after obtaining cross-state results for both NAEP and

TIMSS, I will estimate reduced versions of the model within content topics, and within states, dropping either the topic indicators from the matrix of Item *j* characteristics, or the State *k* principal components scores. To reduce collinearity of the predictors within each of these state-item subpopulations, the raw content emphasis and alignment variables were mean-centered before creating each interaction term.

If the SEC alignment index is functioning well, jointly with cellwise content emphasis, the cellwise curriculum-test alignment underlying the index should predict student performance on corresponding test items. This analysis bears on the validity of the SEC index as a measure of test content representativeness to the extent that instruction is aligned with a state's curriculum and other assumptions, detailed in the next section, hold.

3.3 Assumptions about the US Elementary Education System

Measures of test alignment to a particular curriculum do not address student engagement, pedagogical approaches, or instructional quality (McMaken & Porter, 2012), and their interpretation does not require assumptions about these classroom features, but interpreting results of my models as empirical evidence for validity requires these assumptions. My analysis assumes the average tested student is motivated to engage in learning during classroom instruction, and to solve the NAEP or TIMSS items correctly. One experiment that offered a monetary incentive, awarding examinees $1 for each correct response, to students completing NAEP items found only a small, although statistically significant, effect on mean scores among eighth-graders (O'Neil, Sugrue, & Baker, 1996). To assess how reasonable this assumption is regarding NAEP 2007 scores, I will report descriptive statistics for two variables, by state: students' mean self-reported level of effort on the test, and their feelings about the importance of "succeeding" on NAEP.

Instruction must be aligned with the curriculum in order to produce gains in student achievement on tests that measure curricular objectives (e.g., La Marca et al., 2000). Due to federal school accountability testing in Grades 3–8, which required tests to be aligned to state standards, it would be expected that the taught, or implemented, curriculum closely follows the written, or intended, curriculum, at least relative to magnitudes of alignment typical in previous decades. I assume that state curricula were sufficiently stable over the relevant time period so that teachers might have been familiar with the curriculum objectives, understood them in the manner intended by the state, and have been able to provide instruction targeting them. If instruction targets specific features of state achievement test items (recurring content or formats), rather than curriculum objectives more broadly (e.g., Koretz, 2008), apparently minor differences between the features of state test items and the national assessments' items could influence state mean item performance on the national assessments. To allow cross-state comparability of the results, I assume not only that instruction has followed the curriculum, but also that the degree of instruction-curriculum alignment is similar across states.

For test scores to provide "actionable information" about student learning, test items must be sensitive to instruction; specifically, item difficulty should be a function of exposure to relevant instruction (Mislevy & Zwick, 2012, p. 150). Muthén et al. (1991) showed that the probability of correct response for some eighth-grade mathematics items in the Second International Mathematics Study, particularly items that required definitional knowledge or represented "early stages of learning about selected mathematical topics" (p. 18), depended significantly on whether or not students had received relevant instruction, as reported by their teachers. I will assume that NAEP and TIMSS items are sensitive to differences in content coverage among topics within and across states.

3.4 Assumptions of the Statistical Models

To assess the plausibility of the assumptions of generalized linear models for each item difficulty population (e.g., Breslow, 1996; Gill, 2001), I consider qualitatively a series of diagnostic plots and statistics generated prior to or following estimation of each model. Along with the primary results from regression modeling—model coefficients, effect size measures and model fit statistics—I will report evidence of any serious violation of the models' assumptions, and, when possible, use alternative model specifications to probe the robustness of the results. Scatterplots graphing model predicted and response residual values will be inspected, with any evident patterns in the scatterplot taken to suggest some form of model misspecification: an inappropriate link function or omitted variables (Gill, 2001). Linearity of the relationships between the logit of the expected difficulty values and each predictor, assumed by all models in this study, will be examined by plotting continuous predictor and response residual values for each model, with any patterns of curvature taken to suggest that higher-order predictor terms should be considered for inclusion in the model (Breslow, 1996). To check for evidence of serious collinearity among item- and/or state-specific predictor variable values, prior to each regression analysis, the variables' correlation matrix will be checked, and the variance inflation factor (VIF) value for each predictor will be computed; maximum VIF values exceeding 10 will be reported as a sign that the sample size may be insufficient to obtain precise estimates of some of the regression coefficients.

In analyses, I assume that all sample elements are members of the state-item population of interest. States that had very recently altered their curriculum documents, so that no causal relation between their current curriculum content and item difficulty could be posited, were excluded from the sample. With access to only the released NAEP and TIMSS items, I rely on

the two assessments' frameworks to ensure that the items were relevant to achievement in elementary school mathematics. Observations that have a particularly large influence on the predicted item difficulty values will be identified by computing the Cook's *D* statistic, which quantifies the change in the predicted values produced by deletion of an observation, for each case. Further, observations that have an outsized influence on the effect sizes for the "SEC proportion of curriculum objectives" and "SEC cellwise alignment measure" variables of main interest in this study will be detected by computing approximate DFBETA statistics for each case using linear models of logit-transformed item difficulty values. Characteristics of item cases that have *D* values greater than $4/n$ and/or DFBETA values greater than $2/\sqrt{n}$ (Bollen & Jackman, 1990) will be inspected, and those cases will be evaluated for possible exclusion from the sample. The assessment items are assumed to be representative, if not random, samples from the elementary school mathematics content domains of interest. The state sample is also assumed to be a representative, although not random, sample from the population of US states that had well-established curriculum documents, with each state receiving equal weight; the extent to which this assumption is reasonable will affect the generalizability of the results.

Generalized linear models assume that observations are statistically independent. This assumption may not be plausible for the analysis units in this study, item difficulty values, which are nested within, and likely to exhibit some degree of dependence within, items and states. To quantify the extent to which this assumption is violated by the item difficulty values in each of the two assessment datasets for each grade level, I computed the intraclass correlation coefficient "2,1" from Shrout and Fleiss (1979) within state, and within item, for each dataset. Overall, the intraclass correlation estimates shown in Table 4 indicate very high correlation of item difficulty values within items, as would be expected, and fairly minor correlation of item difficulty values

within states; all of the intraclass correlations were statistically significant at the .05 level. To obtain regression coefficient standard error estimates that are corrected, usually upward, for the effect of the non-independence of observations, a sandwich standard error estimator can be utilized (e.g., Breslow, 1996). However, these standard errors are unbiased only asymptotically as the number of clusters approaches infinity; with small numbers of clusters, they tend to be biased downward, and yield test statistics that do not follow a known distribution (Donald & Lang, 2007). Noting that the number of clusters that should be viewed as "too small" for large-sample inference using the sandwich standard error estimator depends on data features such as disparity in cluster sizes, Cameron and Miller (in press) suggest that 50 clusters may often be inadequate, and 20 clusters will usually be too few. Multilevel modeling, another strategy to account for lack of independence among sample observations, similarly requires a large number of clusters to produce stable results (Raudenbush & Bryk, 2002). In this study, there are only nine or ten states at each grade level, but there are at least 100 items for both assessments at each grade level. Within content topics, the number of items ranges between about 15 and 20, as shown in Tables 1–3. In theory, the item cluster standard error estimator should perform well since the total numbers of items are large, and should produce standard error estimates that are more conservative than the heteroskedasticity-robust standard error estimates, because the proportion of variance in item difficulty that is between items is substantial. The number of states in this study is probably far too small for the state cluster standard error estimator to produce unbiased estimates. To address the non-independence of sample observations, I will compute standard error estimates robust to misspecification of the variance function for the fractional logit model, as recommended by Papke and Wooldridge (1996), standard error estimates adjusted for clustering (non-independence) of the item difficulty outcome variable

values by item, and standard error estimates adjusted for clustering of the item difficulty values by state. I will comment on any differences among these standard error estimates, and report the most conservative standard errors, and *p* values for the corresponding test statistics.

TABLE 4
Intraclass Correlations of Item Difficulty Values within States and Items, by Data Set

	Data Set			
	NAEP Grade 4	NAEP Grade 8	TIMSS Grade 4	TIMSS Grade 8
ICC Within State	0.024	0.038	0.026	0.016
ICC Within Item	0.943	0.927	0.925	0.939

Note. ICC = intraclass correlation
Sources. National Assessment of Educational Progress 2007 (Restricted-Use); Trends in International Mathematics and Science Study 2007.

As an index of the overall explanatory power of each model, the value of an R^2 analog for generalized linear models, the squared correlation between observed and predicted response values, the population value of which equals the average proportion of variance explained by the predictors (Zheng & Agresti, 2000), will be reported, along with its 95% confidence interval. Because predicted values from each model depend on the observed sample values, this R^2 statistic will tend to be biased slightly upward. Zheng and Agresti suggest resampling techniques to create a confidence interval for the R^2, so confidence intervals will be estimated using a nonparametric bias-corrected and accelerated bootstrap method (Efron, 1987). Checking that neither the upper or lower bounds of the confidence interval change by more than .01 when computed from bootstrap procedures using five different seed numbers and several of the models with the smallest sample sizes—within-topic models from the TIMSS Grade 8 data—suggests that 1,000 bootstrap replications will produce reasonably stable estimates of the R^2 confidence interval. For the cross-state models, which have the largest sample sizes, the confidence interval bounds will generally change by less than .003 across 1,000-replication bootstrap procedures using randomly-selected seed numbers.

3.5 Interpretation

I conceive of each cross-state regression analysis using the NAEP or TIMSS data that is interpreted as a primary result addressing Research Question 1 (Grade 4 data) or 3 (Grade 8 data) as essentially a meta-analysis of a collection of independent, equally-weighted within-state studies, using the raw data rather than an effect size measure for each. I will use similar models to replicate my analyses using the two assessments' data for each research question, however. Following analysis, for each of the two sets of results for each research question, I will identify a single preferred model for interpretation, which should typically be the most complete analyzed model—a total of four models for statistical hypothesis testing. For generalized linear models, Stata computes a *z* test statistic that is the square root of the corresponding Wald chi-square statistic for each predictor. In each model, there will be either one or two alignment-index-related predictors of main interest. For Research Question 1, I will be most interested in the coefficient for the proportion of curriculum objectives relevant to a given item's topic and at that item's cognitive demand level. For Research Question 3, I will be primarily interested in the coefficient for the posited interaction between the proportion of the curriculum objectives and the cellwise alignment measure; in the case that the interaction is non-significant, I will plan to judge the significance of the main effects of the proportion of objectives and the alignment measure. Because I will conduct statistical significance testing for two data sets at each grade level, asking similar research questions, it seems that some adjustment of the significance levels for multiple testing is needed, although it is unclear which tests should be identified as constituting a "family" of interrelated tests. Starting from a desired rate of Type I error of .05 for each test on a coefficient of main interest for each research question, I will adjust the significance level for each pair of tests (NAEP and TIMSS data) by research question. Since I

am applying the adjustment within research question, rather than simultaneously across all tests, I will use a Bonferroni correction, which is relatively conservative, as well as easy to implement. Thus, for each of the four tests, I will use a significance level of .025 to judge statistical significance of each variable of main interest as a predictor of test item difficulty. For Research Question 3, if the interaction term in either model is non-significant, I will use a significance level of .0125 to judge the significance of the curriculum objectives proportion and alignment variables. In all models, I will also report whether each predictor reached the three unadjusted benchmark significance levels of .05, .01, and .001, as is conventional, but will not use those results to draw conclusions about the alignment-related predictors of main interest.

Obtaining effect size measures may permit more detailed conclusions about the extent to which results from the four primary models, and the separate within-state and within-topic models described previously, support or fail to support the validity of the SEC alignment measure. Unlike the exponentiated coefficients from a typical logistic regression model with a binary dependent variable, the exponentiated coefficients from a fractional logit model cannot be interpreted as odds ratios. Instead, as an effect size measure, I will report the average marginal effect (AME, or "average partial effect") of each covariate on the outcome, as suggested by Wooldridge (2010, p. 750). The marginal effect of a binary predictor on the expected value of the item difficulty outcome for a particular item-state observation is the difference between the predicted values of the item difficulty, given the unobserved potential and observed values of the binary variable for that observation, the observed values of the other predictors, and the estimated model coefficients. The marginal effect of a continuous predictor on the outcome, for a particular item-state observation, is the partial derivative of the expected value of the outcome with respect to that predictor for that observation. The AME of a predictor, for linear or

nonlinear models, is the mean of the marginal effects, taken over all observations in the sample, and indicates the change in the value of the outcome expected for a one-unit increase in that predictor (Wooldridge, 2010). The AME for a generalized linear model is computed from predicted values given by the inverse link of the linear predictions (which, for fractional logit models, are on the logit scale); thus, for fractional logit models the AME is on the same proportion scale as the observed outcome values. For general linear regression models, the AME for a predictor is the partial regression coefficient. The AME calculation for a predictor in a generalized linear model, such as those used in this study, generalizes the linear regression coefficient to provide an interpretable effect size measure.

As a second type of external validation evidence for the SEC alignment index, addressing Research Question 1, I will examine the relationship between state-level curricular emphasis and instructional emphasis in 2007. To judge the size of the correlation between SEC curricular content emphasis proportions and teacher-reported mean instructional content emphasis, by topic, across states, I will refer to the descriptive terms suggested by Cohen (1992). Because outlying cases may be particularly problematic for the stability of a correlation coefficient estimated from the small sample available for this study, I will also present a scatterplot of the data.

The limited evidence available to support conclusions regarding Research Question 2 suggests that only weak inferences will be possible. Nevertheless, to determine whether the relationship between the proportion of curriculum objectives on a given item's topic at or above the item's cognitive demand level, rather than the proportion of curriculum objectives corresponding to each item's topic-cognitive demand combination, and test item performance should be used as evidence for the validity of the SEC alignment index, addressing Research

Question 2, I will qualitatively compare the relative magnitude of the regression coefficient for the curriculum objectives proportion to that from the corresponding Research Question 1 model, which differs only by substitution of an alternate measure for one predictor—the curriculum proportion. I will also make note of the Bayesian information criterion (BIC) values, an indicator of model plausibility appropriate for comparison of non-nested models like these containing different variants of the curriculum objectives proportion measure.

This chapter described the data, statistical models, and planned analytic strategies to address my three research questions. Necessarily, particular analytic decisions or robustness checks indicated by empirical results from the item difficulty modeling have not been presented here, but will be reported and justified in the next chapter.

CHAPTER 4: RESULTS

This chapter presents primary and ancillary findings regarding the three research questions of interest in this study. The chapter is organized by research question, and contains many results displays. The section for each research question begins with a summary of the findings pertinent to that question; discussion of the findings is reserved for the next chapter. For both the NAEP and TIMSS assessment datasets, model results will first be presented overall, and then by content topic and state for Research Questions 1 and 3. Generally, each table will be presented following its first mention in the text. In keeping with agreements to anonymize alignment findings for some states, states will be identified by generic labels, which will only correspond for the assessments at a particular grade level. Before proceeding to interpret results of the data analyses, I evaluate the extent to which assumptions of the statistical models appear to have been satisfied by the population of state-item observations that were analyzed.

The major threats to unbiased estimation of the regression coefficients in this study were the possibility of important omitted variables potentially correlated with both state-level curricular emphasis and student performance on the standardized test items, the possibility of random measurement error in the predictor variables, and, in the Grade 4 NAEP data, the occurrence of a small fraction of influential cases that caused notable changes in the estimates. Plots of each predictor against the response residual values from each model did not show any nonlinear patterns that would suggest that the functional form of a particular predictor should be reconsidered.

Natural clustering of the item difficulty outcome variable by state and by test item violated the fractional logit models' assumption that errors are independently distributed, yielding standard errors for the regression coefficients that will tend to be underestimated.

Further, given the modest sample sizes used to estimate the item difficulty models and relatively large correlations among the predictors, multicollinearity hindered the ability to separately estimate all coefficients of interest in some of the models. Two types of adjustments to the standard errors to deal with the effects of item and state clustering were entertained. Cluster-robust standard errors based on states as the source of non-independence of observations were considerably smaller than the initial (robust) standard errors for all variables in the cross-state models; because they were consistently more liberal than the baseline estimates, these standard error estimates are not reported. Cluster-robust standard errors with test item as the source of non-independence of observations were larger than heteroskedasticity-robust standard errors for item characteristics variables in the cross-state models, and, in theory, should be preferred over the standard error estimates that treat observations as independent. In results tables for the cross-state models, which are of main interest for statistical hypothesis testing and have a sufficiently large number of distinct item clusters, I will report item-cluster-robust standard errors. In results tables for within-topic or within-state models, since the numbers of item or state groups are quite small, I will report the heteroskedasticity-robust standard errors that treat observations as independent. As explained in the previous chapter, I do not intend to rely on hypothesis testing to interpret the exploratory results from these within-topic and within-state models, so my choice between two standard error estimators that require different, inappropriate assumptions—independent model errors, or a large number of clusters—should not affect my conclusions.

4.1 Research Question 1: Are Counts of Curriculum Objectives a Valid Measure of Curricular Emphasis?

Research Question 1 evaluates two forms of external validation evidence for the SEC alignment index. If proportions of curriculum objectives classified into particular content topic

and cognitive demand domains quantify intended curricular emphasis in a state, and other assumptions about the educational system explicated in Chapter 3 hold, they are expected to be positively correlated with other measures of curricular emphasis beyond a chance level. I first examine the relationship between SEC curricular content analysis proportions and a relatively proximal measure of curricular emphasis: state mean instructional content emphasis ratings by broad mathematics topic, reported by teachers of students in the NAEP sample. These mean instructional content emphasis ratings are taken to represent the implemented curriculum. I then consider associations between SEC curricular content analysis proportions and a set of more distal measures of curricular emphasis: the "attained" curriculum expressed by mean student performance on large-scale mathematics achievement test items.

Results of my analyses indicate a strong, positive, statistically-significant linear relationship between Grade 4 curricular content emphasis proportions and mean instructional emphasis ratings when proportions under broad content topics are consolidated to correspond to the teacher questionnaire categories. Further results suggest that there are statistically significant positive relationships between the average proportion of curriculum objectives corresponding to a particular item's topic-cognitive demand combination in Grades 3 and 4, and classical item difficulty in both the NAEP and TIMSS fourth-grade data; that is, as the proportion of curriculum objectives increases, it is projected that a greater proportion of students will answer the item correctly. However, the size of the average marginal effect is very small: as shown in Model 2 of Tables 5 and 8, only a .013 increase or a .037 increase in NAEP or TIMSS proportion-correct item difficulty values, respectively, would be expected to follow a 10 percentage-point increase in the proportion of curriculum objectives corresponding to a particular item's topic-cognitive demand combination, all else held fixed. Estimation of the

models within content topic reveals that the strength of the relationship between the proportion of objectives measure and item difficulty varies by content topic; for items in some topic areas, there is no apparent relationship between difficulty and the proportion of objectives, although differences in major content categories of the two assessments and limited numbers of items within each make it difficult to compare the NAEP and TIMSS results by topic. Item difficulty models analyzed within states suggest that small positive associations between the proportion of objectives covering a topic and item performance are common across states, and that results of the cross-state significance test are not being driven by unique correlation patterns in one or a few states.

4.1.1 NAEP Grade 4

The first type of validation evidence I considered was the correlation between the cellwise curricular emphasis measures underlying the SEC alignment index and mean instructional coverage ratings from teachers. Teachers of the fourth-grade NAEP examinees were asked about the extent to which their instruction emphasized numbers and operations, measurement, geometry, data analysis, and algebra and functions. Emphasis was reported on a 3-point scale indicating "no," "moderate," or "heavy" emphasis of each content topic. The relationship between the SEC proportion of curriculum objectives measure, condensed across cognitive demand categories, and state mean instructional emphasis by topic is illustrated by the scatterplot in Figure 1. Figure 1 suggests a fairly strong linear relationship between the proportion of curriculum objectives and instructional coverage by topic in the nine states. Both reported Grade 4 instruction and Grade 4 curriculum documents place the greatest emphasis on number properties and basic operations, although the fraction of the curriculum devoted to numbers and operations objectives varies widely by state. Generally, fourth-graders' teachers

report giving heavy emphasis, on average, to number properties and operations, and moderate-to-heavy emphasis to the other four major content strands. They tend to give the least emphasis to data analysis objectives, as appears to be intended by most states' curriculum documents. In most states, students' teachers report giving somewhat greater weight to algebra instruction, on average, than would be projected by the proportion of objectives targeting that topic. The correlation between the state-specific residualized SEC proportion of objectives and the state mean instructional emphasis in nine states was 0.78 ($p < .001$), which is a "large" positive correlation according to Cohen's (1992, p. 157) criteria.

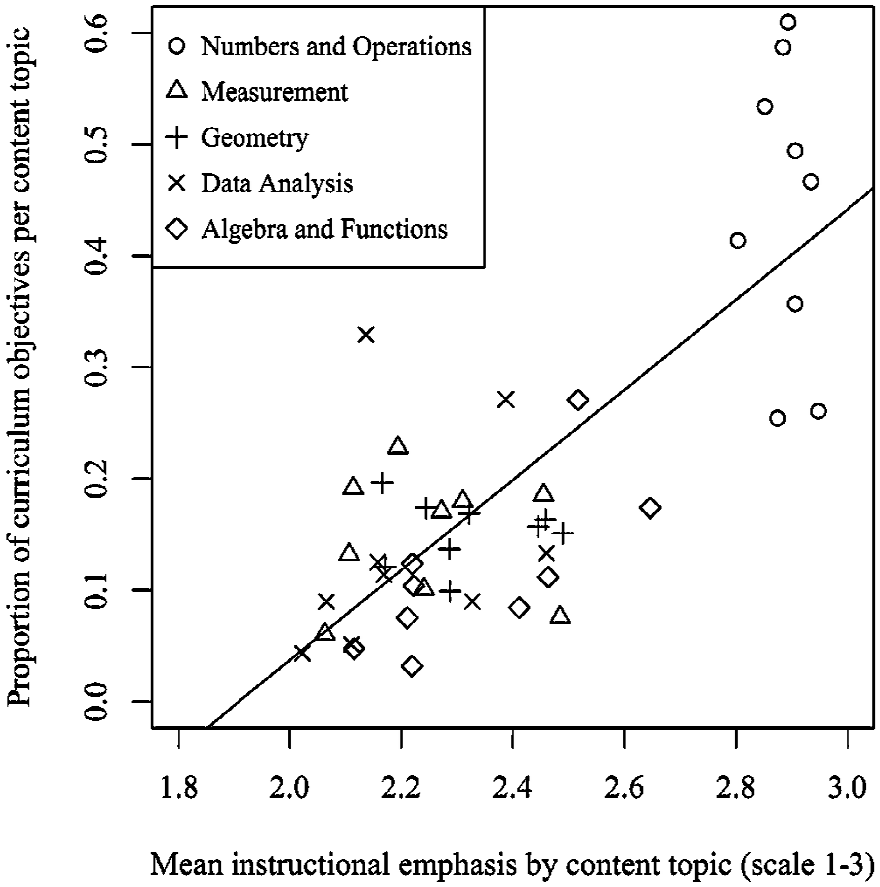

FIGURE 1. Scatterplot of Proportions of Curriculum Objectives and Mean Instructional Emphasis by Mathematics Content Topic for Nine Unidentified States, with Ordinary Least-squares Regression Line.

As a second external source of validation evidence, I modeled the relationship between the cellwise curricular emphasis measures underlying the SEC alignment index and proportion-correct item difficulty values for fourth-graders in nine states on the NAEP and TIMSS 2007 assessments. As shown in Table 5, the results for Model 2, the preferred cross-state model for the Grade 4 NAEP item difficulties, indicate that there is a marginally statistically-significant, but very small, positive relationship (p = .025) between the proportion of curriculum objectives on a particular item's topic and at that item's cognitive demand level in a particular state and proportion-correct item difficulty. Controlling for differences in instructional emphasis and other variables, a 10 percentage-point increase in the proportion of curriculum objectives corresponding to a particular item's topic-cognitive demand combination would be expected to produce only a .013 increase in NAEP item proportion-correct. Comparing the Model 1 and Model 2 estimates for the AME of the proportion of curriculum objectives on item difficulty, displayed in Table 5, it can be noted that the AME is adjusted slightly downward by addition of the teacher-reported instructional emphasis variable to the model, but there is a unique relationship between state-level curricular emphasis and item difficulty that remains even after accounting for differences in instructional emphasis on test items' topics. State mean instructional content emphasis, which each NAEP examinee's math teacher reported on a 4-point scale, is positively related to proportion-correct item difficulty. Item mathematical complexity, low mean state socioeconomic status (principal component 1 scores), and heterogeneity of states' Grade 4 student populations (principal component 2 scores) are negatively related to proportion-correct item difficulty, suggesting that these three control variables are functioning as would be anticipated; directions of coefficients for the other control variables would be difficult to have

predicted in advance. The Model 3 results that appear in Table 5 will be discussed in the section pertaining to Research Question 2.

Sorting DFBETA values by size, I identified one NAEP item that was influential across all states, except Vermont, on the estimated AME of the proportion of curriculum objectives on NAEP item difficulty. This item was a High-complexity Number Properties and Operations item. When the nine state observations on this item was dropped from the cross-state analysis, as shown in appendix Table A4 (which corresponds to the display in Table 5), the AME of the proportion of objectives on NAEP item difficulty was noticeably reduced to .009, although the coefficient was still statistically significant at the .025 level (p = .018). This item was one of the only two High-complexity Number items in the NAEP Grade 4 data. The second High-complexity Number item had a DFBETA value above the cut-off criterion in three states. The relatively high correlations of about .15 between states' proportions of curriculum objectives and their item difficulty on the two High-complexity Number items render the AME larger than it would be absent these items—the estimated AME for the proportion of curriculum objectives is not entirely robust to exclusion of these two items, particularly the one flagged as most influential, from the data. However, although these items are outliers, there is no reason to believe that they should not be considered elements of the population item set. Ideally, it would be good to have more items in this topic-cognitive complexity category to confirm that the AME estimate is not being positively biased by some item feature that is unrelated to the type of academic content targeted by these two items.

Cook's *D* values flagged another NAEP item, a Low-complexity Measurement item, as influential on the slope of the regression across all nine states. This item had unusually low

proportion-correct difficulty values, ranging between .07 and .15, for a Low-complexity item, explaining its outlier status. In spite of the high Cook's *D* values, dropping the nine state

TABLE 5
Fractional Logit Regression Predicting State-Specific NAEP Grade 4 Classical Item Difficulty (*N* = 1458)

	Model 1		Model 2		Model 3	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.061*	0.014	0.057*	0.013		
	(0.025)		(0.025)			
Proportion of curriculum objectives on topic at item cognitive demand level *or higher* (x10)					0.055*	0.013
					(0.024)	
Item topic (*ref* = Number Properties and Operations)						
Measurement	0.107	0.025	0.332	0.077	0.332	0.077
	(0.176)		(0.179)		(0.178)	
Geometry	0.520*	0.119	0.718***	0.164	0.714***	0.163
	(0.205)		(0.206)		(0.205)	
Data Analysis, Statistics, and Probability	0.322	0.075	0.561**	0.129	0.551**	0.127
	(0.212)		(0.210)		(0.209)	
Algebra	0.336	0.078	0.525*	0.121	0.517*	0.119
	(0.222)		(0.220)		(0.219)	
Item complexity (NAEP categories)	-0.852***	-0.199	-0.854***	-0.199	-0.853***	-0.199
	(0.113)		(0.113)		(0.113)	
State principal component 1 score	-0.051***	-0.012	-0.055***	-0.013	-0.054***	-0.013
	(0.003)		(0.003)		(0.003)	
State principal component 2 score	-0.047***	-0.011	-0.045***	-0.01	-0.044***	-0.010
	(0.003)		(0.003)		(0.003)	
State principal component 3 score	-0.001	0.000	-0.009*	-0.002	-0.006	-0.001
	(0.005)		(0.005)		(0.004)	
State principal component 4 score	-0.057***	-0.013	-0.059***	-0.014	-0.058***	-0.013
	(0.004)		(0.004)		(0.004)	
State mean instructional content emphasis on topic (scale 1–3)			0.366***	0.085	0.363***	0.085
			(0.041)		(0.041)	
BIC	-10305		-10298		-10298	
R^2	0.29		0.29		0.29	
R^2 95% CI	0.24, 0.33		0.25, 0.33		0.25, 0.33	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; *ref* = reference group; BIC = Bayesian information criterion; CI = confidence interval

TABLE 6
Fractional Logit Regression Predicting State-Specific NAEP Grade 4 Classical Item Difficulty, by Content Topic

	Number Properties and Operations		Measurement		Geometry		Data Analysis, Statistics, and Probability		Algebra	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.083*	0.019	-0.068	-0.017	0.234	0.050	0.550**	0.129	0.323	0.075
	(0.034)		(0.119)		(0.173)		(0.209)		(0.340)	
Item complexity (NAEP categories)	-0.895***	-0.208	-0.528***	-0.128	-1.220***	-0.259	-0.570***	-0.133	-0.851***	-0.198
	(0.063)		(0.091)		(0.122)		(0.104)		(0.119)	
State mean instructional content emphasis on topic (scale 1–3)	1.967*	0.457	0.233	0.057	0.421	0.089	-0.913	-0.214	-0.240	-0.056
	(0.915)		(0.502)		(0.643)		(0.677)		(0.841)	
State principal component 1 score	-0.055**	-0.013	-0.041	-0.010	-0.043	-0.009	-0.048	-0.011	-0.029	-0.007
	(0.018)		(0.037)		(0.034)		(0.033)		(0.058)	
State principal component 2 score	-0.053**	-0.012	-0.059	-0.014	-0.048	-0.010	-0.029	-0.007	-0.038	-0.009
	(0.019)		(0.035)		(0.035)		(0.035)		(0.038)	
State principal component 3 score	-0.046	-0.011	0.002	0.000	-0.017	-0.004	0.129	0.030	0.052	0.012
	(0.033)		(0.054)		(0.061)		(0.070)		(0.063)	
State principal component 4 score	-0.061	-0.014	-0.023	-0.006	-0.041	-0.009	-0.034	-0.008	0.004	0.001
	(0.038)		(0.079)		(0.076)		(0.073)		(0.086)	
R^2	0.34		0.12		0.46		0.26		0.29	
R^2 95% CI	0.28, 0.41		0.05, 0.18		0.34, 0.56		0.13, 0.40		0.19, 0.40	
N	576		315		207		180		180	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; BIC = Bayesian information criterion; CI = confidence interval

TABLE 7
Fractional Logit Regression Predicting NAEP Grade 4 Classical Item Difficulty, by State (N = 162)

	State A		State B		State C		State D		State E	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.189*	0.043	0.208*	0.048	0.179	0.042	0.210*	0.049	0.310	0.071
	(0.090)		(0.098)		(0.138)		(0.102)		(0.189)	
Item topic (*ref* = Number Properties and Operations)										
Measurement	0.613	0.138	0.748	0.168	0.464	0.106	0.796	0.177	0.757	0.169
	(0.366)		(0.384)		(0.548)		(0.420)		(0.520)	
Geometry	0.939**	0.205	1.064**	0.234	0.868	0.198	1.190**	0.259	1.060*	0.231
	(0.345)		(0.395)		(0.488)		(0.449)		(0.504)	
Data Analysis, Statistics, and Probability	0.886*	0.195	1.045*	0.231	0.661	0.151	1.161*	0.253	0.813	0.181
	(0.384)		(0.457)		(0.559)		(0.539)		(0.442)	
Algebra	0.826*	0.183	1.065*	0.235	0.758	0.173	1.100*	0.241	0.942	0.208
	(0.387)		(0.479)		(0.533)		(0.524)		(0.545)	
Item complexity (NAEP categories)	-0.816***	-0.185	-0.805***	-0.186	-0.830***	-0.196	-0.854***	-0.198	-0.864***	-0.197
	(0.126)		(0.121)		(0.120)		(0.127)		(0.128)	
R^2	0.28		0.29		0.29		0.28		0.31	
R^2 95% CI	0.16, 0.39		0.17, 0.40		0.18, 0.41		0.16, 0.40		0.18, 0.44	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; *ref* = reference group; CI = confidence interval

TABLE 7 (cont'd)

	State F		State G		State H		State I	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.220*	0.052	0.417**	0.097	0.501***	0.119	0.506**	0.119
	(0.107)		(0.154)		(0.135)		(0.175)	
Item topic (*ref* = Number Properties and Operations)								
Measurement	0.625	0.139	0.335	0.077	0.553*	0.115	0.565*	0.130
	(0.346)		(0.211)		(0.216)		(0.239)	
Geometry	1.242**	0.273	1.128**	0.246	1.693***	0.348	1.138***	0.250
	(0.480)		(0.350)		(0.375)		(0.342)	
Data Analysis, Statistics, and Probability	1.130*	0.249	0.831**	0.186	1.795***	0.367	0.623*	0.143
	(0.553)		(0.298)		(0.463)		(0.259)	
Algebra	1.296*	0.284	1.062**	0.233	1.791***	0.366	0.989**	0.221
	(0.580)		(0.397)		(0.479)		(0.370)	
Item complexity (NAEP categories)	-0.792***	-0.188	-0.805***	-0.187	-0.746***	-0.177	-0.708***	-0.166
	(0.123)		(0.129)		(0.124)		(0.128)	
R^2	0.26		0.29		0.27		0.25	
R^2 95% CI	0.14, 0.37		0.17, 0.40		0.16, 0.39		0.13, 0.37	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; *ref* = reference group; CI = confidence interval

observations on this item from the NAEP data and re-estimating the cross-state models had no detectable consequence for the estimated AME of the proportion of curriculum objectives. Thus, this item was retained in the data, although its elimination perhaps could have been justified on the basis of faulty classification.

The overall small positive AME of the proportion of curriculum objectives on item difficulty in Model 2 of Table 5 conceals some heterogeneity in the size of this effect within content topics. The analyses for subsets of the Grade 4 NAEP items by broad mathematics topic, reported in Table 6, suggest that the overall positive effect is driven partly by a large positive AME of the proportion of curriculum objectives on item difficulty for the Data Analysis, Statistics, and Probability items. The model for these items projects a .129 increase in a state's mean item proportion-correct for each .1 increase in the proportion of Data Display, Statistics and Probability-related curriculum objectives that target a given item's cognitive demand level.

Although noting that the coefficient standard errors in the within-state models shown cannot suffer from inflation due to intra-item correlation of item difficulty values as do those in the cross-state models, I will focus on interpreting effect sizes for, rather than statistical significance of, the proportion of curriculum objectives in these models. Multicollinearity among the predictors in these models tended to be high—for some states' models, the maximum VIF value for the predictors exceeded 10—but standard errors for the regression coefficients were not so large as to bar drawing any conclusions from the results. The within-state item difficulty models reported in Table 7 suggest positive relationships between the proportion of curriculum objectives corresponding to a particular item's topic-cognitive demand combination and classical item difficulty. The estimated AME of a 10 percentage-point (i.e., .1) increase in

the proportion of objectives on item difficulty ranges between about .04 and .12 across the nine states.

4.1.2 TIMSS Grade 4

As shown in Table 8, the results for Model 2, the preferred cross-state model for the Grade 4 TIMSS item difficulties, indicate that there is a statistically-significant, but very small, positive relationship (p = .012) between the proportion of curriculum objectives on a particular item's topic and at that item's cognitive demand level and proportion-correct item difficulty in a particular state. All other variables held fixed, a 10 percentage-point increase in the proportion of curriculum objectives corresponding to a particular item's topic-cognitive demand combination would be expected to produce only a .037 increase in TIMSS item proportion-correct. Unlike the NAEP state mean instructional emphasis measure, there is no evidence that the TIMSS state mean percentage of instructional time measure is related to students' average test item performance, but the TIMSS result is based on only two of the nine states in the NAEP sample, and collinearity among predictors in the TIMSS models became particularly acute when state mean instructional time percentage was added, judging from predictors' VIF values, rendering comparison of the NAEP and TIMSS results for this variable difficult. It can further be observed that proportion-correct item difficulty tended to be higher in one of the TIMSS benchmarking states than in the other.

Two TIMSS items had DFBETA values for the proportion of curriculum objectives and Cook's D values that were slightly above their respective cut-off criteria in both states. Both items were Data Display items. One was classed in the Reasoning cognitive domain and had high proportion-correct difficulty values of .93 and .94. The other was a Knowing item with moderate item difficulty values. When the two state observations on either of these items were

dropped from the analysis, the AME for the proportion of curriculum objectives was reduced by .003, as compared to the value shown in Table 8 for the full model, Model 2, but the coefficient was still statistically significant at the .025 level. However, there was little reason to support dropping either of these items from the data. The TIMSS test developers describe their item cognitive domain classification schemes as categorical, not as explicitly ordered—it would be

TABLE 8
Fractional Logit Regression Predicting State-Specific TIMSS Grade 4 Classical Item Difficulty (N = 348)

	Model 1		Model 2		Model 3	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.179*	0.038	0.177*	0.037		
	(0.070)		(0.071)			
Proportion of curriculum objectives on topic at item cognitive demand level *or higher* (x10)					0.165*	0.035
					(0.066)	
Item topic (*ref* = Number)						
Data Display	1.465***	0.264	1.241**	0.226	1.243**	0.226
	(0.327)		(0.433)		(0.435)	
Geometric Shapes and Measures	0.468*	0.102	0.265	0.058	0.259	0.057
	(0.211)		(0.321)		(0.318)	
Item cognitive domain (*ref* = Knowing)						
Applying	-0.217	-0.046	-0.218	-0.046	-0.218	-0.046
	(0.134)		(0.134)		(0.134)	
Reasoning	-0.151	-0.031	-0.159	-0.033	-0.156	-0.032
	(0.337)		(0.340)		(0.340)	
State A	0.233***	0.049	0.215***	0.045	0.234***	0.049
	(0.020)		(0.025)		(0.030)	
State mean percent instructional time on topic			-0.006	-0.001	-0.006	-0.001
			(0.005)		(0.005)	
BIC	-1952		-1946		-1946	
R^2	0.26		0.26		0.26	
R^2 95% CI	0.17, 0.34		0.18, 0.34		0.18, 0.34	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; *ref* = reference group; BIC = Bayesian information criterion; CI = confidence interval

expected, for instance, that on some Knowing items few examinees would respond correctly, and that these items' difficulty values would appear to be regression outliers unless the item sample was very large.

High maximum VIF values for some predictors in the initial TIMSS Grade 4 within-topic models suggested multicollinearity was a serious problem for stability of the coefficients in repeated sampling. To reduce the number of related parameters that had to be estimated, I modified the item cognitive domain variables. On the basis of distributions of the item difficulty values by cognitive domain category, and increasing order of the mean difficulty values for Knowing, Applying, and Reasoning items, I constructed a single variable that treated cognitive domain categories as ordered and linearly related to item difficulty. Replacing the set of cognitive domain indicators in the model with the new cognitive domain variable, plots of residuals against values of this modified predictor indicated that its relationship with item difficulty could reasonably be modeled as linear. Although the separate analyses for Grade 4

TABLE 9
Fractional Logit Regression Predicting State-Specific TIMSS Grade 4 Classical Item Difficulty, by Content Topic

	Data Display		Geometric Shapes and Measures		Number	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.774	0.129	-0.196	-0.043	0.181***	0.039
	(1.105)		(0.123)		(0.041)	
Item cognitive domain (TIMSS categories, ordered)	0.051	0.009	-0.542***	-0.119	-0.045	-0.010
	(0.305)		(0.163)		(0.127)	
State A	0.322	0.054	0.039	0.009	0.262*	0.056
	(0.264)		(0.151)		(0.105)	
R^2	0.04		0.11		0.33	
R^2 95% CI	0.01, 0.23		0.02, 0.25		0.20, 0.45	
N	52		114		182	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; BIC = Bayesian information criterion; CI = confidence interval

TIMSS items by broad mathematics topic, reported in Table 9, have small sample sizes, the model for the Number items fits the data well, judging from the R^2 value and its confidence interval, and estimates a small .039 AME of the curriculum objectives proportion measure on item difficulty that is similar in magnitude to the AME estimated in the overall data.

The within-state item difficulty models reported in Table 10 suggest small positive relationships between the proportion of curriculum objectives corresponding to a particular item's topic-cognitive demand combination and classical item difficulty. The AME of a .1 increase in the proportion of objectives on item proportion-correct ranges is similar in magnitude in the two states, about .03–.04, and also similar to the estimate in the cross-state regression.

TABLE 10
Fractional Logit Regression Predicting TIMSS Grade 4 Classical Item Difficulty, by State (N = 174)

	State A		State B	
	Coef (SE)	AME	Coef (SE)	AME
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.161*	0.033	0.198*	0.043
	(0.068)		(0.077)	
Item topic (*ref* = Number)				
Data Display	1.393***	0.24	1.535***	0.288
	(0.321)		(0.351)	
Geometric Shapes and Measures	0.378	0.081	0.546**	0.122
	(0.222)		(0.210)	
Item cognitive domain (*ref* = Knowing)				
Applying	-0.209	-0.042	-0.226	-0.049
	(0.141)		(0.131)	
Reasoning	-0.165	-0.033	-0.139	-0.03
	(0.339)		(0.354)	
R^2	0.23		0.27	
R^2 95% CI	0.13, 0.34		0.15, 0.38	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; *ref* = reference group; CI = confidence interval

4.2 Research Question 2: Can the Cognitive Demand Categories of the SEC Content Classification Matrix be Treated as Partially Ordered?

The purpose of Research Question 2 is to check the appropriateness of the model underlying Question 1. If cognitive demand is best modeled as an ordinal property of test items, such that instruction requiring application of certain more-demanding cognitive processes related to a particular content topic also benefits students' ability to perform other less-demanding types of cognitive tasks related to the same topic (e.g., Ebel, 1956), models of the relationship between curricular emphasis measures and achievement should account for the proportion of curricular content at or above a particular cognitive level. The follow-up analyses in this study rely on the assumption that the cognitive processes listed by the SEC higher-order cognitive demand categories overlap heavily with those described by particular NAEP mathematical complexity and TIMSS cognitive domain categories.

Revising the proportion of curriculum objectives measure, I find no evidence that substituting the proportion of objectives on a given item's topic at or above the item's cognitive demand level for the proportion of objectives corresponding to each item's topic-cognitive demand combination increases the effect size for the proportion of curriculum objectives, or leads to an improvement in model plausibility as indicated by BIC values. Comparing Models 2 and 3 in Table 5, there is little difference in the estimated AME for the proportion of curriculum objectives on NAEP Mathematics item difficulty, or in overall model fit, whether or not objectives targeting higher levels of cognitive demand related to an item's topic are included in the proportion measure. Table 8 shows that the estimate of the AME of the proportion of curriculum objectives on TIMSS item difficulty is slightly lower when objectives targeting higher levels of cognitive demand related to an item's topic are included in the proportion

measure in Model 3, than when they are excluded from the proportion measure in Model 2, but the BIC indices of overall model fit are identical to the ones place.

These results do not support the conclusions that the SEC cognitive demand categories are semi-ordered and that instruction addressing higher-order cognitive skills benefits performance. However, it should be noted that the proportions of Grade 4 curriculum objectives coded as requiring higher-order reasoning were very small across states, so there was little difference between the curriculum objective proportion measure that was specific to a particular topic-cognitive demand cell, and the measure that counted objectives at the same demand level of higher within a given item's topic. In addition, the proportion of High-complexity items in the NAEP data, which were assumed to require higher-order reasoning, was quite low (see Table 1), so only a small fraction of state-item observations in the NAEP data reflected the change to the proportion of objectives measure.

4.3 Research Question 3: To What Extent are Item-Level Alignment Measures Related to Achievement?

Although I found limited support for the assertion that the proportion of objectives from SEC content analyses is a valid measure of intended curricular emphasis, and predicts achievement on mathematics test items—the main effects of the proportion of objectives measure on Grade 4 item difficulty, while positive and statistically significant, were quite small—in the following analyses I proceed to test the hypothesis that Grade 8 proportion-correct item difficulty will increase with cellwise test-curriculum alignment, at least when curricular emphasis of the material is high. Previous authors have proposed various methods for modeling the posited interaction between alignment and emphasis; in this study, I represent the interaction as a product of the main effects of alignment and emphasis.

Overall, I find little evidence of a relationship between test-curriculum alignment and achievement that depends on curricular emphasis—alignment and curricular emphasis do not appear to interact as hypothesized at Grade 8, at least not after controlling for important covariates such as item complexity. Neither does the average proportion of curriculum objectives measure for Grades 7 and 8 or the cellwise alignment measure appear to be significantly related to test item performance at Grade 8 after controlling for prior topic-specific mean achievement, and other item and state characteristics. Collinearity among the alignment and proportion of objectives measures, their interaction, and the other predictors was considerable in all of the Grade 8 models. Potentially important predictors had to be eliminated from the intended NAEP within-state and within-topic models when near-perfect collinearity prevented their estimation, so those results should be interpreted with particular caution, but the cross-state NAEP and TIMSS models of main interest include the full complement of predictors.

4.3.1 NAEP Grade 8

Results from a fractional logit regression model for NAEP Grade 8 item difficulty that includes only alignment-related predictors, shown in Model 1 of Table 11, depict a positive interaction between the test-curriculum alignment measure and the curricular proportion of objectives measure—predicted item proportion-correct increases with the proportion of objectives when alignment is high, but slightly decreases with the proportion of objectives when alignment is low. Once other covariates are added to the model, however, as shown in the Model 2 results column, the interaction term is no longer statistically significant at the .025 (or even .05) level. Multicollinearity in Model 2 was considerable; the VIF for the NAEP Grade 4 2003 pretest score was greater than 10. Removing the interaction in Model 3, neither the main effect of alignment or of the proportion of objectives is significant.

Although instructional content emphasis information was not collected from eighth-grade teachers, a pretest measure, state mean NAEP Mathematics 2003 Grade 4 scale scores for each content topic, was available for the Grade 8 item difficulty models. States' mean performance on particular content topics during a previous assessment of the 2007 Grade 8 cohort is positively related to proportion-correct item difficulty in this later assessment. As in the Grade 4 NAEP data, item mathematical complexity and low mean state socioeconomic status (principal component 1 scores) are negatively related to proportion-correct item difficulty. Conversely, mean item performance tends to increase with the homogeneity of states' Grade 8 student populations (principal component 2 scores), a relationship that is significant at the .05 level.

One item had high DFBETA values in more than half of the states for both the alignment and proportion of curriculum objectives variables. The item was a High-complexity Algebra item, one of only three High-complexity items administered in Grade 8, all of which were in Algebra (see Table 1). When all ten state observations on this item were dropped from the sample, the estimated AMEs of the proportion of curriculum objectives and alignment measures, which were very small and positive in Model 3 of Table 11, both became slightly negative, but remained very small and non-significant at the .025 level. It seems unsurprising that one of the few High-complexity items, all of which covered the same broad content topic, would appear to be an outlier in the data and would be influential for the full-sample regression estimates, but it is not clear that these estimates should be viewed as biased by inclusion of the state observations on this item. The availability of more High-complexity items would likely have produced more stable regression estimates and permitted more powerful statistical tests of the hypotheses represented by the alignment and proportion of objectives variables in the Table 11 models, but the secondary data analysis presented in this study is limited by the design of the NAEP

TABLE 11
Fractional Logit Regression Predicting State-Specific NAEP Grade 8 Classical Item Difficulty (N = 1670)

	Model 1	Model 2	Model 3	
	Coef (SE)	Coef (SE)	Coef (SE)	AME
Test-curriculum alignment	-2.200*	-0.377	0.040	0.009
	(1.055)	(0.513)	(0.140)	
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.289**	0.061	0.024	0.006
	(0.089)	(0.134)	(0.085)	
Test-curriculum alignment × proportion of curriculum objectives	4.194***	0.932		
	(1.095)	(1.236)		
Topic (*ref* = Number Properties and Operations)				
Measurement		-0.354	-0.392	-0.092
		(0.261)	(0.226)	
Geometry		-0.507*	-0.552**	-0.131
		(0.225)	(0.190)	
Data Analysis, Statistics, and Probability		-0.359	-0.404	-0.095
		(0.297)	(0.258)	
Algebra		-0.286	-0.323	-0.076
		(0.193)	(0.170)	
Item complexity (NAEP categories)		-0.756***	-0.763***	-0.18
		(0.119)	(0.120)	
Mean NAEP Mathematics 2003 Grade 4 subscore		0.011***	0.010*	0.002
		(0.003)	(0.004)	
State principal component 1 score		-0.059***	-0.063***	-0.015
		(0.006)	(0.011)	
State principal component 2 score		0.005	0.011*	0.003
		(0.006)	(0.005)	
State principal component 3 score		-0.002	0.001	0.000
		(0.004)	(0.005)	
State principal component 4 score		-0.038***	-0.044**	-0.01
		(0.008)	(0.015)	
R^2	0.06	0.28	0.28	
R^2 95% CI	0.04, 0.08	0.24, 0.32	0.24, 0.32	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; *ref* = reference group; CI = confidence interval

assessment, which allows for few time-consuming extended reasoning items.

Limiting the analytic sample to only items with a particular content topic, after controlling for other covariates, I found that interactions between the test-curriculum alignment measure and the proportion of objectives measure were not statistically significant for Algebra, Geometry, or Number Properties and Operations items, as presented in Table 12. (Prior to accounting for item mathematical complexity and cohort pretest subscores, there appeared to be a significant positive interaction between the test-curriculum alignment and proportion of objectives measures in predicting Number Properties and Operations item difficulty, and positive main effects of both the alignment and proportion of objectives measures on proportion-correct Algebra item difficulty. Evidently this set of alignment-related predictors had no explanatory value in models for Geometry item difficulty—the lower bound of the 95% confidence interval for the R^2 of the initial model was nearly 0.) Multicollinearity in these models was fairly high, with maximum VIF values exceeding 10 in all three content areas. Removing the interaction term from each model revealed that neither alignment nor the proportion of objectives was a significant predictor of item difficulty within any of the content topics. For the Data Analysis and Measurement item subsets, which had the smallest sample sizes among the topics, models that included both the alignment and proportion of objectives measures encountered estimation problems due to severe multicollinearity; thus, results for these content topics could not be reported.

Similarly, models for state subpopulations of the NAEP Grade 8 data had serious multicollinearity problems. It was not possible to estimate models that included both the alignment and proportion of objectives measures, and indicators for all of the content topic areas.

TABLE 12

Fractional Logit Regression Predicting State-Specific NAEP Grade 8 Classical Item Difficulty, by Content Topic

	Number Properties and Operations				Geometry			
	Coef (SE)	Coef (SE)	Coef (SE)	AME	Coef (SE)	Coef (SE)	Coef (SE)	AME
Test-curriculum alignment	4.477*	1.522	0.663	0.153	-2.181	2.957	0.935	0.224
	(1.822)	(1.890)	(1.070)		(3.460)	(3.418)	(2.420)	
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.652***	0.153	0.035	0.008	0.333	-0.459	-0.099	-0.024
	(0.182)	(0.229)	(0.071)		(0.461)	(0.471)	(0.161)	
Test-curriculum alignment × proportion of curriculum objectives	15.471**	3.405			12.615	-14.771		
	(5.437)	(6.214)			(18.192)	(18.078)		
Item complexity (NAEP categories)		-0.726***	-0.726***	-0.168		-0.771***	-0.771***	-0.184
		(0.070)	(0.070)			(0.097)	(0.097)	
Mean NAEP Mathematics 2003 Grade 4 subscore		0.029**	0.032***	0.008		0.034***	0.031***	0.007
		(0.009)	(0.007)			(0.009)	(0.008)	
R^2	0.04	0.28	0.28		0.00	0.24	0.24	
R^2 95% CI	0.01, 0.09	0.21, 0.37	0.20, 0.36		0.00, 0.03	0.16, 0.33	0.16, 0.33	
N	370	370	370		310	310	310	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; CI = confidence interval

TABLE 12 (cont'd)

	Algebra			
	Coef (SE)	Coef (SE)	Coef (SE)	AME
Test-curriculum alignment	11.296***	0.301	1.818	0.429
	(2.715)	(2.799)	(2.410)	
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	1.057***	0.078	0.162	0.038
	(0.223)	(0.235)	(0.221)	
Test-curriculum alignment × proportion of curriculum objectives	0.145	0.944		
	(0.948)	(0.904)		
Item complexity (NAEP categories)		-0.717***	-0.718***	-0.170
		(0.075)	(0.075)	
Mean NAEP Mathematics 2003 Grade 4 subscore		0.031***	0.029***	0.007
		(0.008)	(0.008)	
R^2	0.11	0.28	0.28	
R^2 95% CI	0.05, 0.18	0.21, 0.38	0.21, 0.37	
N	450	450	450	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; CI = confidence interval

TABLE 13
Fractional Logit Regression Predicting NAEP Grade 8 Classical Item Difficulty, by State (N = 167)

	State J		State K		State L		State M		State N	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Test-curriculum alignment	0.973	0.226	-8.111	-1.913	-10.624	-2.272	-1.216	-0.289	-2.284	-0.542
	(2.885)		(7.610)		(6.965)		(2.84)		(2.392)	
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.182	0.042	0.202	0.048	0.054	0.008	0.214**	0.051	0.275	0.065
	(0.098)		(0.106)		(0.281)		(0.069)		(0.145)	
Test-curriculum alignment × proportion of curriculum objectives					35.952*					
					(16.348)					
Item complexity (NAEP categories)	-0.751***	-0.174	-0.735***	-0.173	-0.760***	-0.177	-0.763***	-0.181	-0.751***	-0.178
	(0.121)		(0.125)		(0.130)		(0.129)		(0.130)	
Mean NAEP Mathematics 2003 Grade 4 subscore	0.009	0.002	0.079	0.019	0.038	0.009	0.002	0.001	-0.018	-0.004
	(0.055)		(0.059)		(0.044)		(0.025)		(0.023)	
R^2	0.24		0.25		0.25		0.26		0.24	
R^2 95% CI	0.14, 0.37		0.14, 0.37		0.14, 0.37		0.15, 0.38		0.13, 0.37	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; CI = confidence interval

TABLE 13 (cont'd)

	State O		State P		State Q		State R		State S	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Test-curriculum alignment	0.100	0.024	-1.416	-0.336	-4.530*	-1.076	-2.082	-0.499	-9.334	-2.234
	(2.078)		(2.805)		(2.216)		(2.127)		(8.872)	
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.153	0.037	0.221	0.053	-0.072	-0.017	0.158	0.038	0.238	0.057
	(0.136)		(0.115)		(0.267)		(0.120)		(0.139)	
Test-curriculum alignment × proportion of curriculum objectives										
Item complexity (NAEP categories)	-0.696***	-0.167	-0.744***	-0.177	-0.804***	-0.191	-0.702***	-0.168	-0.656***	-0.157
	(0.125)		(0.123)		(0.128)		(0.120)		(0.124)	
Mean NAEP Mathematics 2003 Grade 4 subscore	-0.030	-0.007	0.028	0.007	-0.018	-0.004	-0.029	-0.007	0.045	0.011
	(0.028)		(0.022)		(0.024)		(0.018)		(0.043)	
R^2	0.22		0.26		0.26		0.24		0.21	
R^2 95% CI	0.11, 0.34		0.15, 0.38		0.13, 0.37		0.12, 0.36		0.10, 0.33	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; CI = confidence interval

The set of content topic indicators was removed from the model, making omitted variables bias in the regression coefficients shown in Table 13 more likely than if the content topic indicators could have been included. One state, State L, had a statistically significant (at the .05 level) positive interaction between the alignment and proportion of objectives measures after controlling for item mathematical complexity and NAEP Grade 4 pretest scores (p = .028). However, because the actual probability of Type I error for this test was probably considerably greater than .05 due to the number of within-state tests conducted, stability of this result in repeated sampling from that state's population, if such sampling were possible, seems doubtful.

4.3.2 TIMSS Grade 8

As in the Grade 8 NAEP data, regressing only alignment-related predictors on TIMSS Grade 8 item difficulty, as shown in Model 1 of Table 14, there appears to be a positive interaction between the test-curriculum alignment measure and the curricular proportion of objectives measure—predicted item proportion-correct increases with the proportion of objectives when alignment is high, but slightly decreases with the proportion of objectives when alignment is low. Once other covariates are added to the model, however, as shown in the Model 2 results column, the interaction term is no longer statistically significant at the .025 level. Removing the interaction in Model 3, neither the main effect of alignment or of the proportion of objectives is significant at the .0125 level.

To reduce the degree of multicollinearity among predictors in the TIMSS Grade 8 models, which was excessive when separate indicators for each cognitive domain (except one reference domain category) were used as predictors, I constructed a single variable that treated cognitive domain categories as ordered and linearly related to item difficulty. As in the Grade 4

TIMSS data, mean item difficulty values for the cognitive domain categories increased in the order: Knowing, Applying and Reasoning. A plot of the distribution of the item difficulty values

TABLE 14
Fractional Logit Regression Predicting State-Specific TIMSS Grade 8 Classical Item Difficulty (N = 422)

	Model 1	Model 2	Model 3	
	Coef (SE)	Coef (SE)	Coef (SE)	AME
Test-curriculum alignment	-0.821	1.440	1.359	0.312
	(0.882)	(1.026)	(1.075)	
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.046	-0.039	-0.030	-0.007
	(0.052)	(0.090)	(0.082)	
Test-curriculum alignment × proportion of curriculum objectives	5.428***	0.592		
	(1.359)	(1.306)		
Item topic (*ref* = Number)				
Algebra		-0.418*	-0.412*	-0.097
		(0.164)	(0.162)	
Chance		0.306	0.312	0.068
		(0.283)	(0.281)	
Data		0.251	0.271	0.059
		(0.246)	(0.230)	
Geometry		-0.229	-0.227	-0.053
		(0.146)	(0.145)	
Item cognitive domain (TIMSS categories, ordered)		-0.518***	-0.521***	-0.12
		(0.100)	(0.100)	
State J		0.170***	0.167***	0.038
		(0.022)	(0.019)	
R^2	0.05	0.28	0.28	
R^2 95% CI	0.02, 0.10	0.20, 0.34	0.20, 0.34	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; *ref* = reference group; CI = confidence interval

by cognitive domain category, and plots of residuals against values of this predictor, suggested that its relationship with item difficulty could reasonably be modeled as linear. Unlike in the Grade 4 TIMSS results, item cognitive domain was a significant predictor of item difficulty in

many of the Grade 8 models in which it was entered—controlling for other variables, in Model 3 we observe that proportion-correct item difficulty is expected to decrease by .12 (i.e., items becomes more difficult to answer correctly) as cognitive domain is increased to the next ordered category.

Limiting the analytic sample to only items with a particular content topic, after controlling for other covariates, I found that interactions between the test-curriculum alignment measure and the proportion of objectives measure were not statistically significant for any of the content topics, so I removed the interaction term for all of the models. Results for these more restricted models, which appear in Table 15, indicate that neither alignment nor the proportion of objectives is an important predictor of item difficulty within any of the content topics, but the ordered cognitive domain variable again predicts substantial decreases in proportion-correct item difficulty as cognitive domain is increased to the next-most-challenging category in three of the four content topics. It should be noted that the model of interest could not be estimated for the 10 items that covered the "Chance" subtopic in the two states, so these items are not represented in the Table 15 results, although they were included in the overall analyses presented in Table 14.

The results in Table 16 show that, controlling for potential confounding state and item characteristics, the interaction between the test-curriculum alignment measure and the proportion of objectives measure was not statistically significant in either state in the TIMSS Grade 8 data, although the coefficient for the interaction was larger than its standard error in one of the states. Maximum VIF values were greater than 10 in the models that included the interaction and main effects of alignment and the proportion of objectives, and all the covariates. Eliminating the interaction variable from each model, as shown in the final two columns for each state, there was

TABLE 15

Fractional Logit Regression Predicting State-Specific TIMSS Grade 8 Classical Item Difficulty, by Topic

	Algebra		Data		Geometry		Number	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Test-curriculum alignment	-3.670	-0.89	-9.148	-1.895	-0.344	-0.083	4.961	1.094
	(3.398)		(10.076)		(3.897)		(18.333)	
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.136	0.033	-0.282	-0.058	-0.085	-0.021	-0.081	-0.018
	(0.149)		(0.751)		(0.145)		(0.131)	
Item cognitive domain (TIMSS categories, ordered)	-0.128	-0.031	-0.497**	-0.103	-0.551**	-0.134	-0.736***	-0.162
	(0.174)		(0.183)		(0.199)		(0.154)	
State J	0.402	0.098	0.389	0.081	0.175	0.043	0.273	0.06
	(0.265)		(0.333)		(0.155)		(0.870)	
R^2	0.15		0.41		0.17		0.26	
R^2 95% CI	0.06, 0.26		0.21, 0.60		0.05, 0.32		0.15, 0.36	
N	128		60		94		120	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; CI = confidence interval

TABLE 16

Fractional Logit Regression Predicting TIMSS Grade 8 Classical Item Difficulty, by State (N = 211)

	State J				State K			
	Coef (SE)	Coef (SE)	Coef (SE)	AME	Coef (SE)	Coef (SE)	Coef (SE)	AME
Test-curriculum alignment	-6.521**	-3.191	-1.668	-0.379	3.336	4.802	1.743	0.404
	(2.400)	(4.668)	(3.754)		(1.903)	(4.185)	(3.254)	
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.054	-0.064	-0.077	-0.018	-0.114*	-0.195	-0.025	-0.006
	(0.059)	(0.111)	(0.108)		(0.057)	(0.155)	(0.097)	
Test-curriculum alignment × proportion of curriculum objectives	1.095	-2.174			13.091***	9.243		
	(2.459)	(3.378)			(2.270)	(7.191)		
Item cognitive domain (TIMSS categories, ordered)		-0.530***	-0.521***	-0.118		-0.495***	-0.559***	-0.129
		(0.111)	(0.110)			(0.127)	(0.116)	
Item topic (*ref* = Number)								
Algebra		-0.112	-0.227	-0.053		-0.209	-0.424	-0.101
		(0.315)	(0.244)			(0.380)	(0.325)	
Chance		0.354	0.243	0.053		0.607	0.365	0.081
		(0.386)	(0.328)			(0.366)	(0.296)	
Data		0.362	0.269	0.059		0.148	0.323	0.072
		(0.327)	(0.272)			(0.372)	(0.359)	
Geometry		0.014	-0.119	-0.027		0.178	-0.213	-0.05
		(0.305)	(0.198)			(0.380)	(0.197)	
R^2	0.05	0.26	0.26		0.15	0.29	0.28	
R^2 95% CI	0.01, 0.12	0.17, 0.37	0.17, 0.37		0.07, 0.24	0.19, 0.38	0.18, 0.38	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; *ref* = reference group; CI = confidence interval

no statistically significant relationship between the alignment or proportion of objectives measures and item difficulty. Although the estimated AMEs for alignment were quite large, they were in opposite directions in the two states, providing little evidence of a consistent interpretable association between alignment and item difficulty.

4.4 Robustness Check

Although normality of the population error distribution is not an assumption of generalized linear models (Gill, 2001), Breslow (1996) promoted Box-Cox transformation selection as a tool to determine if the link function for a generalized linear model has been misspecified. As noted previously, the exponent of the optimal power transformation to normality for the Grade 8 NAEP item difficulty values, which followed a nearly symmetric but heavy-tailed distribution, determined by the Box-Cox equation was about 1, so the preferred link function identified by Breslow's technique was the identity link. To probe the robustness of the results reported above to potential misspecification of the link in the model for NAEP Grade 8 item difficulty, I used a linear probability model with cluster-robust standard errors to account for high intra-item correlations of the difficulty outcome variable. Estimating the results of models equivalent to those shown in Table 11 using ordinary least squares, I found that the results reported previously for the alignment and proportion of objectives measures were robust to use of a different link function and estimator: when only the alignment and proportion of objectives variables and their interaction were entered in the model, there appeared to be a positive interaction between alignment and the proportion of objectives, but this apparent effect was reduced to non-significance once other predictors were added to the model. Removing the interaction term from the model, the main effects of the alignment and proportion of objectives measures were found not to be statistically significant. Histograms and normal probability plots

of the residuals suggested that the normality assumption was reasonably well-satisfied for all three of these models. The magnitudes of the regression coefficients from these linear models and the corresponding AMEs from the generalized linear models reported in Table 11, which estimate the same quantity in the same metric, were identical to the second decimal place, providing further evidence that the Grade 8 NAEP results are robust to possible misspecification or less-than-ideal use of a logit link for the outcome variable.

CHAPTER 5: CONCLUSIONS AND DISCUSSION

In this final chapter, I draw conclusions regarding each research question, judging the extent to which the results support or fail to support the validity of the SEC alignment index as a measure of test-curriculum correspondence, and compare these conclusions to the findings of previous studies. Evaluating the plausibility of assumptions that were made by the statistical models, in retrospect, I discuss the likely accuracy and stability of the study results as estimates of alignment index component effects in these subpopulations of test items and state curriculum documents, which Cook and Campbell (1979) referred to as "statistical conclusion validity." This discussion establishes the degree to which the model coefficients and standard errors may be considered good estimates for the parameters in the observed population of test items and state curricular content analyses. I then consider the similarity between the theoretical model of the relationships between curricular content emphasis or alignment, instructional emphasis and student achievement, and the statistical model, considering evidence of any departures from the assumptions about the educational system described in Chapter 3. I will discuss the generalizability of the results to other alignment index variants, and other test-state curriculum pairs. The discussions of replicability and generalizability of the results highlight limitations of the study design that qualify the answers to my research questions.

Considering the accumulated evidence, I judge the claim that SEC alignment measures can predict student achievement score gains as still in need of further support. However, this study provided some weak external validation evidence for the curriculum emphasis data: proportions of objectives or "grade-level expectations," used to compute the SEC test-curriculum alignment index. Although the results of this study have limited implications for the selection or use of alignment methods by operational testing programs, evaluated together with results from

previous studies, they suggest the need for two specific lines of future research to bolster claims about the validity of alignment indices as measures of test-curriculum correspondence, and about the validity of conclusions from test-curriculum alignment reviews more generally.

5.1 Research Question 1

The results of this study offer weak support for the validation claim that the proportions of equally-weighted curriculum objectives from SEC content analyses, which underlie the coarse-grained SEC test-curriculum alignment index (Porter, 2002), can be interpreted as measures of intended state curricular emphasis. The detected statistically-significant relationships between a proportion of objectives measure derived from SEC content analysis data, and measures of the implemented and attained Grade 4 mathematics curriculum from nine states were positive, as would be anticipated if proportions of objectives are an accurate measure of curricular emphasis. Models indicate that proportion-correct item difficulty in Grade 4 is expected to increase as the proportion of state curriculum objectives targeting the given item's topic and cognitive demand type in Grades 3 and 4 increases. This conclusion differs from that of a similar study by Mehrens and Phillips (1987). Using textbook emphasis proportions to predict classical item difficulty, they found that textbook emphasis proportion differences, computed after matching Grade 5 and 6 textbook content blocks to a 180-cell classification matrix, had no visible relationship to mathematics achievement test item difficulty (*p*) differences among sixth-graders. However, the effect of a 10-percentage-point increase in the proportion of curriculum objectives corresponding to an item's topic-cognitive demand combination, a sizable shift in the curriculum content, is quite small—between .01 and .03 on average—although the size of the estimated effect varies by content topic. As will be discussed further below, the magnitude of the apparent relationship between the proportion of objectives

and item difficulty may have been attenuated by instruction that deviated from the state curriculum, or inflated by failure to account for important state characteristics that affect both curricular emphasis and test item performance. The proportions of curriculum objectives classified into particular broad content topic domains were also strongly correlated with mean teacher-reported instructional emphasis of those topics in the nine states with available content analyses of established Grade 4 mathematics curriculum documents, even after controlling for state differences in average socioeconomic status and student diversity that might be confounded with distributions of objectives in the curriculum documents. Since all the reported analyses combined proportions of mathematics objectives within a small number of fairly general topics, conclusions about the validity of the fine-grained SEC test-curriculum alignment index do not follow directly from the results of this study. Because no comparison to other alignment indices or formulations of the SEC index were made, I cannot judge whether this index is the best existing measure of intended state curricular emphasis, only that the content analysis proportions that summarize panelists' judgments about objectives' topics and cognitive demand requirements seem to relate in expected ways to other measures of state curricular emphasis in Grade 4.

5.2 Research Question 2

While the preponderance of evidence from previous research on mathematics learning indicates that instruction on higher-order cognitive skills should benefit performance on topical test items that require lower levels of cognitive processing for correct response (e.g., Lobato & Siebert, 2002), modifying the proportion of curriculum objectives measure used in this study, I found no evidence supporting this hypothesis. Accounting for the proportion of objectives at a particular cognitive demand level or higher neither strengthened, nor appreciably weakened, apparent relationships between the curricular proportion measure and item difficulty. Absent

replication with state curricula that show greater variation in coverage of high-complexity objectives, I interpret this result as suggesting (a) insufficient variation in curricular emphasis on highly-demanding objectives, (b) problems with my operationalization of the proportion of objectives measure, and/or (c) violations of my assumptions about the educational system. Although the modified proportion of objectives measure, in principle, could have increased the coverage proportions corresponding to all low- or moderate-demand items (most items in both assessments), because the Grade 3 and 4 state curricula in the sample generally included no, or very few, high-cognitive-demand curriculum objectives, the revised proportion of objectives measure was quite similar to the original measure. High correlation between the modified and original measures likely contributed to the limited change in the estimated relationship between the proportion of objectives and item difficulty when the modified measure was substituted for the original measure in regression models. Also, my empirical model for the relationship between curricular coverage and test item performance assumed that the behaviors defined by the highest NAEP mathematical complexity and TIMSS cognitive domain categories (or least those behaviors that were assessed) would fall into the two highest-demand SEC cognitive demand categories. To the extent that this assumption was violated, differences in model fit when using the original and modified proportion of objectives measures may have been attributable to systematic measurement error. Finally, if instruction in these nine states focused on the portion of the curriculum assessed by state achievement tests as reported by teachers in several states (Stecher et al., 2008), and the state achievement tests seldom assessed the most demanding objectives (e.g., Webb, 1999), no relationship between proportions of high-demand objectives and test item difficulty would be expected; my results might reflect poor correspondence between instruction and the more-demanding segments of the curriculum. While

the results of this study fail to support the hypothesis that the SEC cognitive demand categories are partially ordered in the complexity of cognitive processing required to execute the listed example behaviors in each category, I hesitate to draw any conclusion about the ordinal nature of the categories, and suggest that further empirical and theoretical evaluation is needed to address this question.

5.3 Research Question 3

Despite high collinearity among the predictors in models for Grade 8 item difficulty, my results hinted at positive interactions between cellwise alignment and the proportion of curriculum objectives in Grades 7 and 8. In initial models for NAEP item difficulty by content topic that included only a cellwise alignment measure, a proportion of objectives measure, and their interaction as predictors, the interaction appeared to be an important predictor of Algebra item difficulty, suggesting proportion-correct item difficulty increased with alignment when curricular emphasis (the proportion of objectives) was high. The alignment and proportion of objectives measures also appeared to have positive main effects on Numbers and Operations proportion-correct item difficulty. However, these measures were entirely unrelated to Geometry item difficulty even in this restricted initial model, and once the pretest scores were included in the item difficulty models, there was no statistically-significant interaction between, or unique main effects of, the alignment and proportion of objectives measures in any topic. Results from the TIMSS item difficulty models suggested a potentially important but not statistically-significant interaction between alignment and the proportion of curriculum objectives in one of the two states, but a negative association between alignment and proportion-correct item difficulty in an initial restricted model in the other state. It is possible that the amount of measurement error in the SEC curricular content analysis data, or the fidelity of

instruction to the written curriculum varied considerably in these two states. Because the models of NAEP item difficulty for these two states, which might otherwise be viewed as replications of the TIMSS models, could not incorporate the set of item topic indicators as predictors due to heavy collinearity, they differ meaningfully from the TIMSS models and cannot clarify the reason for the apparent differences in the TIMSS results by state.

Although the relationship between alignment and test item performance at Grade 8 may not be uniform across content topics, the most definitive conclusion of these analyses is that in the presence of other covariates, cellwise test-curriculum alignment measures are not significantly related to test item mathematics performance in Grade 8 on average, even as curricular emphasis on an item's topic and cognitive demand in Grades 7 and 8 increases. This conclusion is consistent with the findings of Mehrens and Phillips (1986) that test-curriculum correspondence ratings were not a significant predictor of school mean mathematics test subscores for third- or sixth-graders, and that differences in classical item difficulty values for schools with high or low test-curriculum correspondence (dichotomizing a 1–5 rating scale) were negligible (Phillips & Mehrens, 1988). This conclusion is also consonant to some degree with the more recent findings of Polikoff and Porter (2012), who tested interactions between an SEC-type instruction-curriculum alignment measure and pedagogical quality measures as predictors of teachers' mean residualized student mathematics achievement scores; while most of the coefficients for interactions predicting mean residualized scores from different mathematics achievement tests were in the positive direction, suggesting teachers' scores increased with alignment when pedagogical quality was relatively high, only 1 out of 12 was significantly different from zero at the .05 level.

The study by Gamoran et al. (1997), which concluded that about one-quarter of the variability in classroom mean mathematics score gains could be explained by a product term computed from instruction-test alignment and reported instructional time measures, is often cited as providing important validation evidence for SEC alignment indices. The alignment measure used in that study was obtained by matching test content to a precursor of the current SEC content analysis matrix. Due to collinearity problems, the research model did not include main effects of instruction-test alignment or instructional emphasis together with the product term that otherwise would have been interpretable as an interaction. Using more typical linear regression models including main effects of instruction-curriculum or test-curriculum alignment and content emphasis, as well as their product, as predictors of classroom or state mean achievement, Polikoff and Porter's recent (2012) study and this study have failed to convincingly reproduce Gamoran et al.'s findings. Operationalizing instruction-test alignment and instructional emphasis measures using ratings from two different researcher-developed instruments, however, D'Agostino et al. (2007) found that the interaction between alignment and emphasis was a positive, significant predictor of fifth-graders' math scores in a multilevel model, lending some support to Gamoran et al.'s conclusion that instruction-test alignment influences test score gains, at least when instructional emphasis on the tested material is high.

A simple interpretation of the mixed results could be that instruction-test alignment matters for test performance gains, but the additional assumptions about the educational system required to also link instruction-curriculum or test-curriculum alignment to achievement outcomes (e.g., the outcome measure corresponds to the curriculum, instruction follows the curriculum) do not hold in practice. Other explanations for the difficulty replicating Gamoran et al.'s findings using SEC-type alignment measures may be that the relatively high-poverty

schools sampled by the study were not sufficiently representative of the general school population (Schmidt & Maier, 2009), or that the instruction-curriculum alignment variable used captured differences in pedagogical quality or mean student affluence that were not represented elsewhere in the math achievement model. A final possible explanation for the inconsistent conclusions about the empirical relationship between SEC alignment indices and achievement is that the Gamoran et al. research model tested the hypothesis that mean mathematics achievement score gains increased as content coverage—either alignment or instructional time, or both increased. Although that study's results have been interpreted as demonstrating the effect of curricular alignment on achievement test scores, the reported positive effect of the alignment-by-emphasis product term on score gains may have been driven primarily by instructional time on spent on tested topics (e.g., Coates, 2003), rather than by instruction-test alignment. Considering the accumulated research, overall, I judge the claim that SEC alignment measures can predict student achievement score gains as still in need of further support.

5.4 Accuracy of the Results in the Mathematics Achievement Test Item-State Population

The previous discussion presenting the conclusions of this study and comparing them to others' findings implicated model design differences that may have contributed to discrepancies in our conclusions about the validity of SEC-type alignment indices. Limitations of this study arising from failure of my assumptions for statistical estimation and hypothesis testing, or mismatch between my theoretical and statistical models of the data population also could have caused my results to diverge from those of previous studies. I will comment first on the most doubtful assumptions underlying my statistical conclusions. To the extent that the assumptions of the fractional logit models used in this study are not satisfied in the test item-state population, the computed regression coefficients or their standard errors could under- or overestimate the

true population parameters. Lack of independence of item-state observations in the cross-state models was addressed through use of a sandwich standard error estimator, but, as in many studies based on observational data, the possibility of important omitted predictors or systematic measurement error in the predictors remains a threat to the accuracy of my statistical conclusions.

Interpreting the regression coefficients from item difficulty models as unbiased estimates of population quantities requires the assumption that no relevant variables have been omitted from the models. Failure to control for state or item characteristics that co-vary with states' curricular emphasis patterns *and* cause differences in state item difficulty could result in biased estimates of the effects of alignment, the proportion of curriculum objectives, or their interaction, on item difficulty. In this study, the omitted variable of greatest concern may be average state mathematics achievement in each content topic at the beginning of Grade 3, before students were exposed to instruction on the Grade 3 and 4 math curricula. The models for Grade 4 item difficulty that were used to address Research Questions 1 and 2 did not account for possible existing state differences in average, topic-specific mathematics ability at the start of Grade 3 because large-scale achievement testing of students prior to Grade 3 seldom occurs, and so no suitable measure of initial state mean mathematics topic achievement was available. Because state-level mathematics curriculum development before 2006 was not always methodical, it is not immediately clear that there would be a relationship between students' mean subtopic performance early in Grade 3, and the distribution of objectives in the Grade 3 and 4 curriculum documents. If there were such a relationship, it is impossible to predict what its direction might be, and so difficult to anticipate the direction of any bias in the alignment-related regression coefficients that might result. The models for Grade 8 NAEP item difficulty that were used to

address Research Question 3 did include a topic-specific mathematics pretest score for each state, although it was measured at the end of Grade 4, rather than at the beginning of Grade 7 instruction; lack of a well-timed pretest measure may have again resulted in some degree of bias to the regression coefficients unless achievement differences among states in each topic remained approximately constant between the end of Grade 4 and start of Grade 7. Although ideal mathematics pretest measures for each state were not available, all models accounted for state differences in child poverty rates that could be expected to serve, to some degree, as a proxy for initial mean mathematics achievement.

Previous studies of mathematics achievement test item difficulty indicate that the most important predictors of difficulty are items' mathematical complexity and linguistic features (Enright & Sheehan, 2002; Shaftel et al., 2006). All models controlled for item mathematical demands. While I did not have access to information about item linguistic features, linguistic features would not be expected to have any relation to state mathematics curricular emphasis patterns, unless incidentally, so inclusion of these relevant item characteristics in the models would not be expected to change the size of regression coefficients for the alignment-related variables, although it might adjust their standard errors. It is also difficult to conceive of the proportion of curriculum objectives measure acting as a proxy for some unobserved item characteristic, such that the apparent effect of curricular emphasis on item difficulty in the Grade 4 data is actually due to unmeasured features of the test items.

As well as requiring the assumption of no omitted variables, interpreting the regression coefficients and average marginal effects from fractional logit models as unbiased estimates necessitates the assumption that predictor variables are measured without error. In this study, the variables likely to contain the most random measurement error are the predictors of main

interest: the alignment and proportion of curriculum objectives measures, and the teacher instructional emphasis ratings in the Grade 4 data. Consolidating objective proportions to the coarse-grained level should reduce the fraction of random measurement error in the proportion measure, which has already been averaged over multiple judges, relative to that which would be found in proportion measures derived from fine-grained SEC content analysis based on a larger classification matrix (Mehrens & Phillips, 1986). However, if the fine-grained rather than the coarse-grained objective proportion and alignment measures are theorized to be related to student achievement outcomes (Porter, 2002), then the coarse-grained variables used in this study should be viewed as fine-grained variables measured with error, and their relationship to achievement should be interpreted as likely attenuated relative to the relationship that would be expected if fine-grained alignment-related variables could be linked to the NAEP and TIMSS test items. The coarse-grained proportion of objectives measure, which averaged cellwise curriculum proportions across the grades during and immediately prior to testing, also ignored any effects of curriculum coverage in earlier grades (subsequent to the Grade 4 pretest measure, in the Grade 8 NAEP data) on achievement, perhaps further weakening the observed relationship between curricular emphasis and state item difficulty. The instructional emphasis ratings self-reported by teachers, used as a control variable in some Grade 4 model specifications, are also likely to contain measurement error. If teachers systematically over-report their emphasis on all curriculum topics, only the model intercepts will be biased—not a major concern. If, however, there is substantial random measurement error in the state mean instructional emphasis ratings, which are correlated with the curricular emphasis proportion measure, bias in the regression coefficient for the curriculum proportion measure could result. The estimated regression coefficients for curricular emphasis in the full models that include instructional emphasis may be

too large. I would argue, though, that any effect of instructional emphasis on achievement could be considered a part of the effect of curricular emphasis on achievement, with curricular emphasis as the *a priori* cause (e.g., Holland, 1986), so that the reduced models excluding the possibly unreliable instructional emphasis variable may produce the best estimates for the relationship between the proportion of curriculum objectives and item difficulty. Average marginal effects for the proportion of objectives measures in these reduced models for the Grade 4 NAEP and TIMSS data, reported in Chapter 4, were small, positive, and very similar in magnitude to those interpreted in my conclusions.

5.5 Defensibility of Assumptions about the US Elementary Education System

If a test-curriculum alignment index is determined not to be positively related to student achievement gains as theorized, or to be only weakly related to achievement gains, it may be that the index is based on invalid measures of content emphasis, or otherwise not functioning as a meaningful quantitative variable. However, numerous alternative explanations are possible (Porter, 2006). My theoretical model of mathematics item difficulty for elementary school students in the US essentially combined the model of curricular achievement from the early TIMSS studies (Travers & Westbury, 1989) with recent models of item difficulty on state mathematics achievement tests (Ferrara et al., 2011; Shaftel et al., 2006). Serious departure of my theoretical model and assumptions from realities of the elementary education system in the mid-2000s could lead to incorrect conclusions about the practical importance of curricular emphasis and test-curriculum alignment for mathematics test item performance, even if the statistical conclusions presented previously are accurate. In particular, relationships between alignment-related measures and item difficulty in the large-scale mathematics assessment data, which appeared to be weakly positive (Grade 4) or null (Grade 8), could have been attenuated to

the extent that instruction did not follow the state curricula, test items were not sensitive to instruction, or student motivation during test-taking was low. This section will weigh each of these counter-explanations.

The high correlation between curricular content emphasis proportions and mean teacher-reported instructional emphasis at the broad topic level observed in this study suggests that instruction, on average, did not diverge widely from the content specified by state curricula, but this correlation reflects a macro-level view of instruction-curriculum correspondence, and does not imply any direction of causality. Elementary mathematics teachers in these states may have willfully decided not to follow the specific lists of curriculum objectives given in curriculum documents, interpreted curriculum objective statements differently than state policy makers intended (e.g., Spillane, 2004), or, perhaps most likely, diverged from the curriculum due to reliance on state- or district-adopted instructional materials (e.g., Senk & Thompson, 2003) that were not designed to match their state's curriculum. However, as described in Chapter 2, teachers were under significant pressure to deliver instruction following their state curriculum, and, in surveys of three states between 2004 and 2006 (Stecher et al., 2008), most elementary and middle school math teachers reported having modified the content of their instruction to better address state curriculum objectives, so there is some evidence that teachers were aware of state curriculum documents and intended their instruction to target grade-level student performance goals.

Achievement test items are often assumed or explicitly claimed to be sensitive to instruction following a particular curriculum. Although the items used in this study were not designed to test attainment of objectives set forth in a single specific curriculum document, both TIMSS and NAEP are intended as tests of school mathematics achievement (Mullis et al., 2005;

NAGB, 2006), rather than, for instance, mathematics literacy like the Programme for International Student Assessment studies. Instructional sensitivity of many items in IEA mathematics assessments prior to TIMSS 2007 has been demonstrated (Miller & Linn, 1988; Muthén, 1988; Schmidt et al., 1999). Because the TIMSS 2007 assessment framework draws heavily on the frameworks of previous IEA studies, performance on the TIMSS items would be expected to be influenced by differences in instruction across educational jurisdictions. While the NAEP assessment framework and test development procedures differ from those of TIMSS, Muthén et al. (1991) asserted that instructional sensitivity appeared to be highest for definitional and other low-complexity items, which comprise the largest fraction of the NAEP 2007 items, suggesting that many NAEP items should also be sensitive to differences in examinees' instructional histories. The positive relationships generally observed in this study between the proportion of curriculum objectives measure and classical item difficulty indicate that the response processes for these assessments' items can be influenced by state differences in instruction, as posited by Grissmer et al. (2000) regarding NAEP scale scores.

To gauge whether low student motivation during test-taking was likely to have distorted the relationship between alignment and test item performance, I examined self-reports of the NAEP examinees on two questionnaire items about effort. Among Grade 4 examinees who responded to the effort questions in the nine states that had SEC curricular content analyses, tabulated in appendix Table A5, between one-tenth and one-fifth viewed success on NAEP as "somewhat" or "not" important, rather than "important" or "very important." Similar proportions of fourth-graders asserted that they had not tried as hard on NAEP as on other tests. There was no obvious relationship between the proportions of fourth-graders responding affirmatively to these two prompts in the nine states. Eighth-graders were more likely than

fourth-graders to make a distinction between the perceived importance of performing well on NAEP, and the level of effort they actually exerted in responding to the test items. While between one-third and more than half of Grade 8 students, depending on state, recognized NAEP as a low-stakes test, only about one-fifth of them asserted that they had offered less than complete effort on the assessment. Many eighth-graders may have perceived some value in the test-taking experience besides tangible rewards or sanctions, which they recognized would be limited (e.g., Brophy & Ames, 2005). Judging from their self-reports, eighth-graders in Alabama and Kansas appear to have been more likely than those in the other eight states, on average, to have put forth at least as much effort in completing the NAEP items as they would in taking other tests, while those in Massachusetts seem to have been less likely to devote full effort to NAEP test-taking. Comparing the rightmost columns of appendix Table A5 in the eight states that were analyzed at both grade levels, overall, higher proportions of Grade 8 than Grade 4 students reported exerting less than their full effort on the NAEP assessment, a pattern of decreasing effort that would be expected across the NAEP-tested grade levels (Brophy & Ames, 2005). However, even among eighth-graders, the hypothesized relationship between alignment and mean test item performance appears unlikely to have been much attenuated by lack of examinee motivation on the outcome tests, presuming that students engaged similarly with the TIMSS assessment tasks.

5.6 Generalizability of the Results to Other Alignment Indices and State Curriculum Documents

As suggested by the discussion above, the findings of a weak positive relationship between the "coarse-grained" (Porter, 2002) SEC proportion of curriculum objectives and item difficulty in the Grade 4 data, and no relationship between curricular emphasis or test-curriculum

alignment measures and item difficulty in the Grade 8 data, would be expected to have limited generalizability to "fine-grained" versions of the same measures. Porter (2002, 2006) has contended that the alignment among curriculum documents, instruction, and assessments that matters most for student achievement is alignment at the level of specific content performance goals—matching of emphasis or instructional time at the level of broad content strands is argued to be inadequate to produce well-aligned tests, or achievement gains, depending on the intended use of an alignment index. The associations between analogous measures constructed from fine-grained SEC content analysis matrices and mathematics item difficulty or instructional content emphasis could be either larger or smaller in this state-item sample than those reported for the coarse-grained measures. The results of this study do not have direct implications for the validity of SEC test-curriculum alignment indices in other content areas (e.g., English, Language Arts, and Reading), or of SEC-type indices based on different content classification matrices (e.g., Liu & Fulmer, 2008).

While my conclusions strictly deal with test-curriculum alignment indices only, validity evidence for test-curriculum alignment indices may bear on the functioning and interpretation (Porter et al., 2007) of instruction-curriculum alignment indices generated from SEC data by teachers, administrators and researchers. Because the validation evidence collected for the SEC instruction-curriculum index thus far does not provide consistent support for the index as a measure of instruction-curriculum correspondence when controlling for preexisting group achievement differences (Gamoran et al., 1997; Polikoff & Porter, 2012), and this study produced only limited support for the curriculum emphasis proportions underlying SEC instruction-curriculum alignment indices, the need to further evaluate the validity of these indices is suggested.

The findings of this study cannot be readily generalized to support the validity of results from other popular alignment methods (e.g., Webb, Achieve, Human Resources Research Organization) because the types of results generated by various methods (e.g., qualitative or quantitative, single criterion or multiple criteria) differ considerably. While conclusions regarding the appropriateness of using unweighted counts of curriculum objectives to compute test-curriculum alignment indices would be equally relevant to the Webb balance-of-representation index and the SEC alignment index of interest in this study, the small magnitude of the positive relationships between SEC curricular emphasis proportions and Grade 4 test item difficulty observed in this study is not clearly attributable to problems with the objective weighting scheme. Based on this and previous research, the SEC index might be judged to be better-supported by external validation evidence than other test-curriculum alignment indices, simply because no comparable information for other methods or indices has been published, although rater agreement data are now typically reported for various alignment methods' document reviews.

The external validation evidence for the coarse-grained SEC test-curriculum alignment index produced by this study would be expected to have some generalizability to the group of states with established mathematics curriculum documents during this time frame. The weak or absent connection observed between alignment-related measures and NAEP mathematics item difficulty, based on test item performance in eleven states from different regions of the US, would be projected to have more generalizability to the US populations of Grade 4 and 8 students in 2007 than the TIMSS results, which reflect test item responses in only two states. Although elements and overall emphases of state curriculum documents at the same grade level are likely to have varied widely in 2007 (Porter et al., 2009; Reys et al., 2007), they were more

similar in format than during previous decades (e.g., Webb, 2007), so relationships between the document content and other indicators of curricular emphasis could be predicted to be similar to those observed in this study's cross-state sample. Appendix Table A6, which displays means of state means for selected characteristics in the study states and all 50 states in Grades 4 and 8, suggests that the state curriculum content analysis samples should not be claimed to be representative of those from all states, as the study states, for instance, have a lower mean proportion of students eligible for the federal school meal program and higher average mean NAEP Mathematics scale scores than observed in all states at both grade levels. Even if the study states appeared to be representative with respect to these mean characteristics, they may not have been representative with respect to curriculum emphases. The results are more likely to generalize to states that had longstanding curriculum documents than to states with curriculum documents that were under initial or re-development.

5.7 Suggestions for Future Validation Research

This study sought to investigate the validity of the interpretation of a commonly-used alignment index as a measure of test-curriculum correspondence by examining linear relationships between components of the index and concurrent measures of state curricular emphasis in mathematics. I interpreted the results as providing weak external validation evidence for the content analysis summary data—proportions of objectives or "grade-level expectations,"— which are used to compute the SEC test-curriculum alignment index, as curriculum emphasis measures. However, even taken together with other validation studies' results, there is little compelling support for any existing alignment method's rating data as a replicable, meaningful indicator of test-curriculum correspondence. Additional scrutiny of commonly-used alignment methods appears to be warranted. Noting that alignment

methodologies themselves cannot be considered "valid," Davis-Becker and Buckendahl (2013, p. 24) recommend collecting several types of evidence to validate the inferences from results of a given alignment study: procedural evidence (e.g., rater qualifications, execution of rater training), internal evidence (e.g., rater agreement, reliability coefficients), external evidence (e.g., replication studies), and utility evidence (i.e., observed usefulness of results to test and curriculum developers). Of the four types of evidence, internal agreement measures would seem to be the most easily reported, but also could be inflated by implementation of consensus-seeking steps within a particular alignment procedure. Although not always systematically documented, the procedural information they list could usually be recorded in a straightforward manner following established quality-monitoring procedures for standard-setting studies (e.g., Cizek & Bunch, 2007). External validation evidence, though, may be expensive and thus difficult to collect (Davis-Becker & Buckendahl). Utility judgments would appear to be relevant only after alignment results have otherwise been established as sound.

Kane (2013) argues that validation studies should concentrate on testing the most doubtful claims in a particular test score interpretive argument. For the alignment indices and other alignment results used in validation of achievement test score interpretations, the most crucial, yet dubious claims may be that (1) overall and item-level alignment conclusions are replicable across independent panels trained by different facilitators, and (2) the content classification schemes used capture distinct types of performances with mathematical content. There is a need for evidence showing that alignment matching or rating frameworks have been systematically developed, reviewed and potentially revised by diverse groups of curriculum and rater cognition experts. Likewise, there is a need to demonstrate that alignment results can be sufficiently replicated across independent review occasions, using real data collection as well as

generalizability projections or agreement indices estimated from single panels. Since reproducibility will depend on features of the content classification scheme, these two claims are interrelated. The additional claim that proportions of curriculum objectives corresponding to particular content strands indicate intended emphasis by state policymakers seems also doubtful, but perhaps less fundamental than claims about the meaningfulness of the classification scheme used and reproducibility of results.

Advising that the organizations sponsoring particular test-curriculum alignment reviews (often state education agencies) will seldom have the resources to finance collection of external validation evidence, Davis-Becker and Buckendahl note that assessment professionals have the responsibility to report and interpret such evidence if it is available. They suggest that external validation evidence could include evaluations of the same test-curriculum pair by alignment panels using multiple methods or the same method, results from other types of content analysis studies, or comparisons with test item content classifications assigned by item writers. Thus far, the primary external validation claim made for SEC alignment indices has been that they predict student achievement gains (e.g., Porter, Polikoff, Barghaus, & Yang, 2013), so the present study sought evidence relevant to this claim. However, interpreting alignment indices, or indeed any alignment results, as meaningful indicators of test-curriculum correspondence also minimally requires a claim that the results would not vary too greatly if an independent panel of qualified experts (having adequate familiarity with the given curriculum and examinee population), trained by a different facilitator, conducted the content analysis. This claim is implicit in interpretation of results from any alignment review, including but not limited to SEC procedures, and is necessary to warrant further claims about state achievement test scores, for instance, that they measure students' mastery of curriculum objectives. While monitoring curricular alignment and

demonstrating its impact on student learning gains may be separately of interest, the most direct external validation evidence to obtain for results of SEC and other alignment methods, requiring no assumptions about instructional quality, would be that the results are reproducible. Such a validation study would require greater resources than computation of agreement or generalizability coefficients, but could demonstrate that alignment results are not heavily dependent on the particular panelists selected (Webb et al., 2007) or on anomalies in the training or matching process, but rather represent an interpretation of the degree of test-curriculum correspondence that would be largely held in common by qualified experts.

The procedural validation evidence outlined by Davis-Becker and Buckendahl (2013) implies that high-quality, high-fidelity implementation of any existing alignment method that evaluates both content topic and cognitive demand match between tests and curriculum documents *may* yield valid conclusions regarding test-curriculum alignment. However, not all test content classification schemes will be of equal quality or utility (Schmidt & Maier, 2009). Recent studies have particularly questioned the extent to which expert judges can reliably categorize test item content using published cognitive demand coding schemes (Schneider et al., 2013). In addition to seeking empirical evidence of alignment index validity as in the present study (see also, e.g., Porter et al., 2008; Webb et al., 2005), future development of commonly-used alignment indices should call on sizable, diverse groups of subject-matter curriculum experts and learning scientists to evaluate the theoretical underpinnings of the classification schemes used to rate behavioral tasks from tests and curriculum documents (Schmidt & Maier, 2009); topic or cognitive demand categories that are viewed as ill-defined, overlapping, or inconsistent with knowledge of mathematics learning should be revised. Future research should consider whether content topics are too fine-grained, or not fine-grained enough (Porter et al.,

2013), and whether the descriptions, labels and/or examples used define cognitive demand categories are sufficiently distinct from one another, and comprehensive of the cognitive processing or behaviors that could be required by achievement test items. Consideration of cognitive demand classification schemes in mathematics is particularly needed to find schemes that can be reliably utilized by item raters (e.g., Ferrara et al., 2011) and item writers (e.g., Porter et al., 2013), and ideally are also sensible from a cognitive science perspective. Alternatively, if devising cognitive demand schemes that can be consistently applied by alignment panelists and item writers proves infeasible, it may be necessary, for curricular achievement tests, to simply deem test items judged to require the particular behaviors described by curriculum objectives as "aligned" and items requiring other behaviors, even those that may require similarly-difficult cognitive processing, as not aligned (e.g., D'Agostino et al., 2008). This method would highlight current measurement limitations of large-scale testing, but could provide a realistic, transparent assessment of test-curriculum match. Regardless of the test and curriculum content classification scheme adopted for a particular alignment review, detailed information about its development process, and about the consistency with which it can be applied by independent content experts on different occasions, should be viewed as important pieces of validation evidence in interpretive arguments for state achievement test scores as measures of curricular attainment.

APPENDIX

TABLE A1

SEC Task Cognitive Demand

Category	Description
Memorize	Illustrative examples: Recite basic mathematical facts. Recall mathematics terms and definitions. Recall formulas and computational procedures.
Perform procedures	Illustrative examples: Use numbers to count, order or denote. Perform computational procedures or algorithms. Follow procedures/instructions. Make measurements. Solve equations or routine word problems. Organize or display data. Read or produce graphs and tables. Execute geometric constructions.
Demonstrate understanding	Illustrative examples: Communicate mathematical ideas. Use representations to model mathematical ideas. Explain findings and results from data analysis. Develop/explain relationships between concepts. Explain relationships between models, diagrams or other representations.
Conjecture, generalize, prove	Illustrative examples: Determine the truth of a mathematical pattern or proposition. Write formal or informal proofs. Analyze data. Find a mathematical rule to generate a pattern or number sequence. Identify faulty arguments or misrepresentations of data. Reason inductively or deductively. Use spatial reasoning.
Solve non-routine problems, make connections	Illustrative examples: Apply and adapt a variety of appropriate strategies to solve problems. Apply mathematics in contexts outside of mathematics. Synthesize content and ideas from several sources.

Note. From CCSSO & WCER (2004).

TABLE A2

NAEP Item Mathematical Complexity

Category	Description
Low complexity	This category relies heavily on the recall and recognition of previously learned concepts and principles. Items typically specify what the student is to do, which is often to carry out some procedure that can be performed mechanically. It is not left to the student to come up with an original method or solution. The following are some, but not all, of the demands that items in the low-complexity category might make: Recall or recognize a fact, term, or property. Recognize an example of a concept. Compute a sum, difference, product, or quotient. Recognize an equivalent representation. Perform a specified procedure. Evaluate an expression in an equation or formula for a given variable. Solve a one-step word problem. Draw or measure simple geometric figures. Retrieve information from a graph, table, or figure.
Moderate complexity	Items in the moderate-complexity category involve more flexibility of thinking and choice among alternatives than do those in the low-complexity category. They require a response that goes beyond the habitual, is not specified, and ordinarily has more than a single step. The student is expected to decide what to do, using informal methods of reasoning and problem-solving strategies, and to bring together skill and knowledge from various domains. The following illustrate some of the demands that items of moderate complexity might make: Represent a situation mathematically in more than one way. Select and use different representations, depending on situation and purpose. Solve a word problem requiring multiple steps. Compare figures or statements. Provide a justification for steps in a solution process. Interpret a visual representation. Extend a pattern. Retrieve information from a graph, table, or figure and use it to solve a problem requiring multiple steps. Formulate a routine problem, given data and conditions. Interpret a simple argument.

TABLE A2 (cont'd)

High complexity	High-complexity items make heavy demands on students, who must engage in more abstract reasoning, planning, analysis, judgment, and creative thought. A satisfactory response to the item requires that the student think in abstract and sophisticated ways. Items at the level of high complexity may ask the student to do any of the following: Describe how different representations can be used for different purposes. Perform a procedure having multiple steps and multiple decision points. Analyze similarities and differences between procedures and concepts. Generalize a pattern. Formulate an original problem, given a situation. Solve a novel problem. Solve a problem in more than one way. Explain and justify a solution to a problem. Describe, compare, and contrast solution methods. Formulate a mathematical model for a complex situation. Analyze the assumptions made in a mathematical model. Analyze or produce a deductive argument. Provide a mathematical justification.

Note. From U.S. Department of Education, National Assessment Governing Board (2006, pp. 36–40).

TABLE A3

TIMSS Item Cognitive Domain

Category	Description
Knowing	Knowing covers the facts, procedures, and concepts students need to know. This cognitive domain covers the following behaviors: Recall definitions, terminology, number properties, geometric properties, and notation. Recognize mathematical objects, shapes, numbers and expressions, and mathematical entities that are equivalent. Carry out algorithms for addition, subtraction, multiplication or division, or a combination of these, with whole numbers, fractions, decimals or integers. Approximate numbers to estimate computations. Carry out routine algebraic procedures. Retrieve information from graphs, tables or other sources. Read simple scales. Using measuring instruments, use units of measurement appropriately and estimate measures. Classify/group objects, shapes, numbers or expressions; make correct decisions about class membership. Order numbers and objects by attributes.
Applying	Applying focuses on the ability of students to apply knowledge and conceptual understanding to solve problems or answer questions. This cognitive domain covers the following behaviors: Select an efficient/appropriate operation, method or strategy for solving problems where there is a known algorithm or method of solution. Display mathematical information in diagrams, tables, charts or graphs, and generate equivalent representations for a given mathematical entity or relationship. Generate an appropriate model, such as an equation or diagram, for solving a routine problem. Follow and execute a set of mathematical instructions. Given specifications, draw figures or shapes. Solve routine problems, similar to those encountered in class (e.g., use geometric properties to solve problems; compare and match different data representations [Grade 8]; use data from charts, graphs or maps).

TABLE A3 (cont'd)

Reasoning	Reasoning goes beyond the solution of routine problems to encompass unfamiliar situations, complex contexts, and multi-step problems. This cognitive domain covers the following behaviors:
	Determine and describe or use relationships between variables or objects in mathematical situations. Use proportional reasoning (Grade 4). Decompose geometric figures to simplify solving a problem. Draw the net of a given unfamiliar solid. Visualize transformations of three-dimensional figures. Compare and match different representations of the same data (Grade 4). Make valid inferences from given information.
	Extend the domain to which the results of mathematical thinking and problem solving are applicable by restating results in more general, widely applicable terms.
	Combine mathematical procedures to establish results, and combine results to produce a further result. Make connections between different elements of knowledge and related representations, and link related mathematical ideas.
	Provide a justification for the truth or falsity of a statement by reference to mathematical results or properties.
	Solve non-routine, unfamiliar problems in mathematical, real-life and/or complex contexts. Use geometric properties to solve non-routine problems.

Note. From Mullis et al. (2005, pp. 33–38).

TABLE A4

Fractional Logit Regression Predicting State-Specific NAEP Grade 4 Classical Item Difficulty, Dropping One Influential Item (N = 1449)

	Model 1		Model 2		Model 3	
	Coef (SE)	AME	Coef (SE)	AME	Coef (SE)	AME
Proportion of curriculum objectives on topic at item cognitive demand level (x10)	0.039*	0.009	0.035*	0.008		
	(0.016)		(0.017)			
Proportion of curriculum objectives on topic at item cognitive demand level *or higher* (x10)					0.035*	0.008
					(0.016)	
Item topic (*ref* = Number Properties and Operations)						
Measurement	0.043	0.010	0.272	0.064	0.275	0.064
	(0.171)		(0.174)		(0.174)	
Geometry	0.445*	0.103	0.647**	0.148	0.648**	0.149
	(0.200)		(0.202)		(0.202)	
Data Analysis, Statistics, and Probability	0.242	0.057	0.485*	0.113	0.483*	0.112
	(0.204)		(0.203)		(0.203)	
Algebra	0.248	0.058	0.440*	0.102	0.440*	0.102
	(0.213)		(0.213)		(0.214)	
Item complexity (NAEP categories)	-0.839***	-0.196	-0.840***	-0.197	-0.840***	-0.196
	(0.115)		(0.115)		(0.115)	
State principal component 1 score	-0.050***	-0.012	-0.054***	-0.013	-0.053***	-0.012
	(0.003)		(0.003)		(0.003)	
State principal component 2 score	-0.046***	-0.011	-0.044***	-0.01	-0.044***	-0.01
	(0.003)		(0.003)		(0.003)	
State principal component 3 score	0.002	0.001	-0.006	-0.001	-0.004	-0.001
	(0.004)		(0.004)		(0.003)	
State principal component 4 score	-0.056***	-0.013	-0.058***	-0.014	-0.058***	-0.013
	(0.004)		(0.004)		(0.004)	
State mean instructional content emphasis on topic (scale 1–3)			0.375***	0.088	0.373***	0.087
			(0.039)		(0.039)	
R^2	0.269		0.271		0.271	

Notes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; AME = average marginal effect; *ref* = reference group; BIC = Bayesian information criterion

TABLE A5

Measures of Average Test-taking Effort by NAEP 2007 Examinees, by Grade and State

	Proportion reporting success on NAEP is somewhat or not important				Proportion reporting not trying as hard on NAEP as on other tests			
	Grade 4		Grade 8		Grade 4		Grade 8	
	p	*SE*	*p*	*SE*	*p*	*SE*	*p*	*SE*
AL			0.36	0.009			0.18	0.008
CA	0.14	0.004			0.21	0.005		
IN	0.15	0.007	0.49	0.010	0.16	0.007	0.20	0.008
KS	0.10	0.006	0.39	0.010	0.11	0.006	0.14	0.007
MA	0.17	0.007	0.60	0.010	0.11	0.006	0.26	0.009
MI	0.16	0.007	0.48	0.010	0.18	0.007	0.21	0.009
MN	0.16	0.007	0.52	0.010	0.12	0.006	0.21	0.008
NJ			0.54	0.010			0.25	0.008
OH	0.13	0.006	0.50	0.010	0.15	0.007	0.22	0.009
OR	0.18	0.007	0.54	0.010	0.16	0.007	0.22	0.009
VT	0.18	0.008	0.57	0.011	0.10	0.006	0.23	0.010

Source. National Assessment of Educational Progress 2007.

TABLE A6

Average Means of Selected State Characteristics for Study and All States in 2007, by Grade

	Grade 4				Grade 8			
	Study		All		Study		All	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Proportion minority students	0.291	0.060	0.351	0.027	0.260	0.033	0.334	0.026
Proportion students in rural schools	0.336	0.069	0.384	0.028	0.384	0.062	0.401	0.027
Proportion federal school meal program-eligible students	0.392	0.028	0.438	0.016	0.344	0.022	0.390	0.015
Mean NAEP Mathematics scale score	243	2.60	240	0.899	286	2.01	281	0.898

Source. National Assessment of Educational Progress 2007.

REFERENCES

REFERENCES

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.

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

American Federation of Teachers. (2008). *Sizing up state standards 2008*. Washington, DC: Author.

Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory Into Practice, 41*(4), 255–260.

Beck, M. D. (2007). Review and other views: Alignment as a psychometric issue. *Applied Measurement in Education, 20*(1), 127–135.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Fredriksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359). Hillsdale, NJ: Erlbaum.

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

Blank, R. K., & Smithson, J. (2009). *Alignment content analysis of TIMSS and PISA Mathematics and Science assessments using the Surveys of Enacted Curriculum methodology*. Paper prepared for the National Center for Education Statistics and American Institutes for Research.

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

Bollen, K. A., & Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 257–291). Newbury Park, CA: Sage.

Breslow N. E. (1996). Generalized linear models: Checking assumptions and strengthening conclusions. *Statistica Applicata, 8*, 23–41.

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

Brophy, J., & Ames, C. (2005). *NAEP testing for twelfth graders: Motivational issues*. Washington, DC: National Assessment Governing Board. Retrieved from ERIC database. (ED500959)

Brown, R. S., & Conley, D. T. (2007). Comparing state high school assessments to standards for success in entry-level university courses. *Educational Assessment, 12*(2), 137–160.

Cameron, A. C., & Miller, D. L. (in press). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*. Retrieved from http://cameron.econ.ucdavis.edu/research/papers.html

Carmichael, S. B., Martino, G., Porter-Magee, K., & Wilson, W. S. (2010). *The state of state standards—and the Common Core—in 2010*. Washington, DC: Thomas B. Fordham Institute.

Chalifour, C., & Powers, D. E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement, 26*(2), 120–132.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Coates, D. (2003). Education production functions using instructional time as an input. *Education Economics, 11*(3), 273–292.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Council of Chief State School Officers & Wisconsin Center for Educational Research [CCSSO & WCER]. (2004, October 25). *Coding procedures for curriculum content analyses*. Retrieved from https://secure.wceruw.org/seconline/Reference/CntCodingProcedures.pdf

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). New York: Chapman and Hall.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Crocker, L., Llabre, M., & Miller, M. D. (1988). The generalizability of content validity ratings. *Journal of Educational Measurement*, *25*(4), 287–299.

Crocker, L. M., Miller, M. D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education*, *2*(2), 179–194.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

D'Agostino, J. V., Welsh, M. E., Cimetta, A. D., Falco, L. D., Smith, S., VanWinkle, W. H., & Powers, S. J. (2008). The rating and matching item-objective alignment methods. *Applied Measurement in Education, 21*(1), 1–21.

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessments. *Educational Assessment*, *12*(1), 1–22.

D'Agostino, R. B., Belanger, A., & D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *American Statistician*, *44*(4), 316–321.

Davis-Becker, S. L., & Buckendahl, C. W. (2013). A proposed framework for evaluating alignment studies. *Educational Measurement: Issues and Practice*, *32*(1), 23–33.

Donald, S. G. & Lang, K. (2007). Inference with differences-in-differences and other panel data. *Review of Economics and Statistics*, *89*(2), 221–233.

Ebel, R. L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement*, *16*(3), 269–282.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychological Science Quarterly*, *50*(3), 328–344.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129–157). Mahweh, NJ: Erlbaum.

Ferrara, S., & Duncan, T. (2011). Comparing science achievement constructs: Targeted and achieved. *The Educational Forum*, *75*, 143–156.

Ferrara, S., Duncan, T., Freed, R., Velez-Paschke, A., McGivern, J., Mushlin, S., . . . Westphalen, K. (2004, April). Examining test score validity by examining item construct validity: Evidence of the alignment of observed skills, cognitive processes, and response strategies with test specifications. In E. A. Vanderputten (Chair), *Putting alignment to the test*. Symposium conducted at the Annual Meeting of the American Educational Research Association, San Diego, CA. Retrieved from http://www.air.org/files/AERA2004MidSchlScienceAssess.pdf

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test design with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice, 30*(4), 3–15.

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 225–243). New York: Springer.

Floden, R. (2002). The measurement of opportunity to learn. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 231–266). Washington, DC: National Academies Press.

Flowers, C., Wakeman, S., Browder, D. M., & Karvonen, M. (2009). Links for Academic Learning (LAL): A conceptual model for investigating alignment of alternate assessments based on alternate achievement standards. *Educational Measurement: Issues and Practice, 28*(1), 25–37.

Freeman, D. J., Belli, G. M., Porter, A. C., Floden, R. E., Schmidt, W. H., & Schwille, J. R. (1983). The influence of different styles of textbook use on instructional validity of standardized tests. *Journal of Educational Measurement, 20*(3), 259–270.

Frisbie, D. A. (2003). *Checking the alignment of an assessment tool and a set of content standards*. Iowa Technical Adequacy Project (ITAP). Iowa City, IA: University of Iowa.

Fulmer, G. W. (2011). Estimating critical values for strength of alignment among curriculum, assessments, and instruction. *Journal of Educational and Behavioral Statistics, 36*(3), 381–402.

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis, 19*(4), 325–338.

Gill, J. (2001). *Generalized linear models: A unified approach*. Thousand Oaks, CA: Sage.

Glatthorn, A. A. (1999). Curriculum alignment revisited. *Journal of Curriculum and Supervision, 15*(1), 26–34.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35.

Grissmer, D. W., Flanagan, A., Kawata, J. H., & Williamson, S. (2000). Improving student achievement: What state NAEP test scores tell us. Santa Monica, CA: RAND.

Guion, R. E. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, *1*(1), 1–10.

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, *5*(10), 511–517.

Haertel, E. A. (1985). Content validity and criterion-referenced testing. *Review of Educational Research*, *55*(1), 23–46.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333.

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore: Johns Hopkins University Press.

Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Mahwah, NJ: Erlbaum.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 47–76). New York: Routledge.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education*, *24*, 393–446.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessment: A case study. *Applied Measurement in Education*, *20*(1), 101–126.

Hill, H. C. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal*, *38*(2), 289–318.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Jones, D. H., & Szatrowski, T. H. (1983). On the statistical determination of content validity. *Educational and Psychological Measurement, 43*(4), 995–1004.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 39–64). Charlotte, NC: Information Age.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Klein, D. (2005). *The state of state math standards 2005*. Washington, DC: Thomas B. Fordham Institute.

Klein, S. P., & Kosecoff, J. P. (1975). Determining how well a test measures your objectives. Los Angeles: Center for the Study of Evaluation, University of California. Retrieved from ERIC database. (ED109226)

Koretz, D. (2008). Further steps toward the development of accountability-oriented science of measurement. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 71–91). New York: Taylor & Francis.

Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2010). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *Journal of Special Education, 44*(3), 131–145.

La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation, 7*(2).

La Marca, P. M., Redfield, D., & Winter, P. (2000). *State standards and state assessment systems: A guide to alignment*. Washington, DC: Council of Chief State School Officers.

Leighton, J. P., & Gokiert, R. J. (2008). Identifying potential test item misalignment using student verbal reports. *Educational Assessment, 13*(4), 215–242.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18*(2), 85–96.

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics, 21*(1), 83–90.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. *Applied Psychological Measurement, 4*(4), 547–561.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of the requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3–16.

Liu, X., & Fulmer, G. W. (2008). Alignment between science curriculum and assessments in selected New York State Regents exams. *Journal of Science Education and Technology, 17*(4), 373–383.

Lobato, J., & Siebert, D. (2002). Quantitative reasoning in a reconceived view of transfer. *Journal of Mathematical Behavior, 21*(1), 87–116.

Martineau, J., Paek, P., Keene, J., & Hirsch, T. (2007). Integrated, comprehensive alignment as a foundation for measuring student progress. *Educational Measurement: Issues and Practice, 26*(1), 28–35.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218.

Martone, A. (2007). *Exploring the impact of teachers' involvement in an assessment-standards alignment study*. Unpublished doctoral dissertation, University of Massachusetts Amherst.

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

McMaken, J., & Porter, A. (2012). The Surveys of Enacted Curriculum as a measure of implementation. In D. J. Heck, K. B. Chval, I. R. Weiss, & S. W. Ziebarth (Eds.), *Approaches to studying the enacted mathematics curriculum* (pp. 173–193). Charlotte, NC: Information Age.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*(3), 185–196.

Mehrens, W. A., & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357–370.

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

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*(3), 205–219.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). New York: American Council on Education and MacMillan.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 83–108). Charlotte, NC: Information Age.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 97–128). Mahweh, NJ: Erlbaum.

Mislevy, R. J., & Zwick, R. (2012). Scaling, linking, and reporting in a periodic assessment system. *Journal of Educational Measurement, 49*(2), 148–166.

Mullis, I. V., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: Boston College, Lynch School of Education, TIMSS & PIRLS International Study Center.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks*. Chestnut Hill, MA: Boston College, Lynch School of Education, TIMSS & PIRLS International Study Center.

Muthén, B. O. (1988). *Instructionally sensitive psychometrics: Applications to the Second International Mathematics Study* (CSE Technical Report 286). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.

Muthén, B. O., Kao, C.-F., & Burstein, L. (1991). Instructionally-sensitive psychometrics: Application of a new IRT-based detection technique to mathematics test items. *Journal of Educational Measurement, 28*(1), 1–22.

Neidorf, T.S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 assessments* (NCES 2006-029). Washington, DC: US Department of Education, National Center for Education Statistics.

No Child Left Behind Act of 2001. Pub. L. No. 107-110 U. S. C. §115, Stat. 1450 (2002).

Notice of Final Priorities for Race to the Top Fund, 74 Fed. Reg. 59,688 (Nov. 18, 2009).

Olson, J. F., Martin, M. O., & Mullis, I. V. (Eds.). (2008). *TIMSS 2007 Technical Report*. Washington, DC: US Department of Education, National Center for Education Statistics.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. L. (1996). Effects of motivational interventions on NAEP mathematics performance. *Educational Assessment, 3*(2), 135–157.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics, 11*(6), 619–632.

Peak, H. (1953). Problems of objective observation. In L. Festinger & D. Katz (Eds.), Research methods in the behavioral sciences (pp. 243–300). New York: Dryden Press.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy. (ED481836)

Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 733–755). Westport, CT: American Council on Education and Praeger.

Phillips, S. E., & Mehrens, W. A. (1988). Effects of curricular differences on achievement test data at item and objective levels. *Applied Measurement in Education, 1*(1), 33–51.

Plake, B. S., Impara, J. C., & Buckendahl, C. W. (2004). Technical quality criteria for evaluating district assessment portfolios used in the Nebraska STARS. *Educational Measurement: Issues and Practice, 23*(2), 12–16.

Poggio, J. P., Glasnapp, D. R., Miller, M. D., Tollefson, N., & Burry, J. A. (1986). Strategies for validating teacher certification tests. *Educational Measurement: Issues and Practice, 5*(2), 18–25.

Polikoff, M. S. (2012a). Instructional alignment under No Child Left Behind. *American Journal of Education*, *118*(3), 341–368.

Polikoff, M. S. (2012b). The association of state policy attributes with teachers' instructional alignment. *Educational Evaluation and Policy Analysis, 34*(3), 278–294.

Polikoff, M. S., & Fulmer, G. W. (2013). Refining methods for estimating critical values for an alignment index. *Journal of Research on Educational Effectiveness, 6*(4), 380–395.

Polikoff, M. S., & Porter, A. C. (2012). *Surveys of Enacted Curriculum Substudy of the Measures of Effective Teaching Project: Final report*. Retrieved from http://www.aefpweb.org/sites/default/files/webform/FINAL%20SEC%20REPORT_Polikoff_Porter.pdf

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

Porter, A. C. (2006). Curriculum assessment. In J. Green, G. Camilli, & P. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141–160). Washington, DC: American Educational Research Association.

Porter, A. C., McMaken, J., & Blank, R. K. (2011). Surveys of Enacted Curriculum and the State School Officers Collaborative. In W. F. Tate, K. D. King, & C. R. Anderson (Eds.), *Disrupting tradition: Research and practice pathways in mathematics education* (pp. 21–31). Reston, VA: National Council of Teachers of Mathematics.

Porter, A. C., Polikoff, M. S., Zeidner, T., & Smithson, J. (2008). The quality of content analyses of state student achievement tests and state content standards. *Educational Measurement: Issues and Practice*, *27*(4), 2–14.

Porter, A. C., Polikoff, M. S., Barghaus, K. M., & Yang, R. (2013). Constructing aligned assessments using automated test construction. *Educational Researcher*, *42*(8), 415–423.

Porter, A. C., Polikoff, M. S., & Smithson, J. (2009). Is there a de facto national intended curriculum? Evidence from state content standards. *Educational Evaluation and Policy Analysis*, *31*(3), 238–268.

Rabinowitz, S., Roeber, E., Schroeder, C., & Scheinker, J. (2006, January). *Creating aligned standards and assessment systems* (Issue paper 3). Retrieved from http://www.ccsso.org/Documents/2006/Creating_Aligned_Standards_2006.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Reckase, M. D., & Chen, J. (2012). The role, format, and impact of feedback to standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 149–164). New York: Routledge.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2004). Benchmarking and alignment of standards and testing. *Educational Assessment*, *9*(1&2), 1–27.

Reys, B., Chval, K., Dingman, S., McNaught, M., Regis, T. P., & Togashi, J. (2007). Grade-level learning expectations: A new challenge for elementary mathematics teachers. *Teaching Children Mathematics*, *14*(1), 6–11.

Roach, A. T., McGrath, D., Wixson, C., & Talapatra, D. (2010). Aligning an early childhood assessment to state kindergarten content standards: Application of a nationally recognized alignment framework. *Educational Measurement: Issues and Practice*, *29*(1), 25–37.

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools*, *45*(2), 158–176.

Robitaille, D., Schmidt, W. H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *The Third International Mathematics and Science Study: Curriculum frameworks for mathematics and science* (Monograph No. 1). Vancouver, Canada: Pacific Educational Press.

Rothman, R. (2003, March). *Imperfect matches: The alignment of standards and tests*. Paper commissioned by the National Research Council, Center for Education, Committee on Test Design for K-12 Science Achievement. Washington, DC: National Academy of Sciences.

Sanford, E. E., & Fabrizio, L. M. (1999, April). *Results from the North Carolina-NAEP comparison and what they mean to the End-of-Grade Testing Program*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Schafer W. D., Wang, J., & Wang, V. (2009). Validity in action: State validity evidence for compliance with NCLB. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 173–193). Charlotte, NC: Information Age.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: American Council on Education/Praeger.

Schmidt, W. H., & Maier, A. (2009). Opportunity to learn. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 541–559). New York: Routledge.

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (1999). *Facing the consequences: Using TIMSS for a closer look at U.S. mathematics and science education*. Dordrecht, The Netherlands: Kluwer.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning.* San Francisco, CA: Jossey-Bass.

Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment, 18*(2), 99–121.

Senk, S. L., & Thompson, D. R. (2003). *Standards-based school mathematics curricula: What are they? What do students learn?* Mahweh, NJ: Erlbaum.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105–126.

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

Sireci, S. G., & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16*(1), 17–31.

Smithson, J. L., & Collares, A. C. (2007, April). Alignment as a predictor of student achievement gains. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Snow, R. E. (1994). A person-situation interaction theory of intelligence in outline. In A. Demetriou & A. Efklides (Eds.), *Intelligence, mind, and reasoning: Structure and development* (pp. 11–28). Amsterdam: North-Holland.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: Macmillan.

Spillane, J. S. (2004). *Standards deviation: How schools misunderstand education policy*. Cambridge, MA: Harvard.

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., . . . Naftel, S. (2008). *Pain and gain: Implementing No Child Left Behind in three states, 2004–2006.* Santa Monica, CA: RAND.

Swanson, C. B., & Stevenson, D. L. (2002). Standard-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis, 24*(1), 1–27.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*(4), 901–926.

Travers, K. J., & Westbury, I. (1989). *The IEA Study of Mathematics I: Analysis of mathematics curricula*. Oxford: Pergamon.

Truxillo, C. (2005). Maximum likelihood parameter estimation with incomplete data. *Proceedings of the Thirtieth Annual SAS(r) Users Group International Conference*. Retrieved from: http://www2.sas.com/proceedings/sugi30/111-30.pdf

Turner, L. C., & Carlson, L. (2003). Indexes of item-objective congruence for multidimensional items. *International Journal of Testing, 3*(2), 163–171.

University of Wisconsin, Wisconsin Center for Educational Research, Measures of Enacted Curriculum Group [MECG]. (2004, October 25). *Coding procedures for curriculum content analyses*. Retrieved from https://secure.wceruw.org/seconline/Reference/CntCodingProcedures.pdf

University of Wisconsin, Wisconsin Center for Educational Research, Measures of Enacted Curriculum Group [MECG]. (2010, November 9). *Mathematics content analysis* [Data files]. Retrieved from http://seconline.wceruw.org/MSP/Content/ELA/ELACntRpt/WSELACntRptMenu.asp

US Department of Education. (2004, April). *Standards and assessments peer review guidance: Information and examples for meeting the requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author. Retrieved from http://dese.mo.gov/divimprove/fedprog/grantmgmnt/NCLB_PDF/Standards_Assessemnts_Peer_Review_Guidance_04282004.pdf

US Department of Education. (2012, June). *ESEA flexibility: Review guidance for Window 3*. Washington, DC: Author. Retrieved from Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html\

US Department of Education, Institute of Education Sciences, National Center for Education Statistics [NCES]. (2008). *Digest of education statistics* (2008 ed.) [Statistical tables]. Retrieved from http://nces.ed.gov/programs/digest/2008menu_tables.asp

US Department of Education, Institute of Education Sciences, National Center for Educational Statistics [NCES] (2009, May 13). *NAEP technical documentation*. Retrieved from http://nces.ed.gov/nationsreportcard/tdw/

US Department of Education, Institute of Education Sciences, National Center for Educational Statistics [NCES] (2013). *NAEP data explorer* [Statistical tables]. Retrieved from http://nces.ed.gov/nationsreportcard/naepdata/dataset.aspx

US Department of Education, National Assessment Governing Board [NAGB]. (2006, Sept.). *Mathematics framework for the 2007 National Assessment of Educational Progress*. Washington, DC: Author.

Vockley, M. (2009). *Alignment and the states: Three approaches to aligning the National Assessment of Educational Progress with state assessments, other assessments, and standards*. Washington, DC: Chief Council of State School Officers.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. *Research Monograph No. 8*. Washington, D.C.: Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. *Research Monograph No. 18*. Madison, WI: University of Wisconsin, National Institute for Science Education.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7–25.

Webb, N. L., Alt, M., Ely, R., Cormier, M., & Vesperman, B. (2005). *The Web Alignment Tool: Development, refinement, and dissemination*. Washington, DC: Council of Chief State School Officers.

Webb, N. L., Alt, M., Ely, R., & Vesperman, B. (2005). Web Alignment Tool (WAT): Training Manual, Version 1.1. Retrieved from http://wat.wceruw.org/index.aspx

Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice, 26*(2), 17–29.

Welsh, M. E., D'Agostino, J. V., & Kaniskan, B. (2013). Grading as a reform effort: Do standards-based grades converge with test scores? *Educational Measurement: Issues and Practice, 32*(2), 26–36.

Wiley, D. E., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis, 17*(3), 355–370.

Winfield, L. F. (1993, April). *Investigating test content and curriculum content overlap to assess opportunity to learn*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Wixson, K. K., & Yochum, N. (2004). Research on literacy policy and professional development: National, state, district, and teacher contexts. *The Elementary School Journal, 105*(2), 219–242.

Woolard, J. C. (2007, April). *Measuring systemic alignment of a state's instruction, standards, and assessments: A baseline analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed). Cambridge, MA: MIT Press.

Wyse, A. E., & Viger, S. G. (2011). How item writers understand depth of knowledge. *Educational Assessment, 16*(4), 185–206.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12*(8), 10–14.

Zheng, B., & Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine, 19*(13), 1771–1781.