

**Validity of Comparing Test Scores on State Assessments with the Results of the National Achievement of Educational Progress**

**Liru Zhang**  
**Delaware Department of Education**  
[lzhang@doe.k12.de.us](mailto:lzhang@doe.k12.de.us)

**Paper presented at the 2008 AERA Annual Conference**  
**New York City, NY, March, 2008**

# Validity of Comparing Test Scores on State Assessments with the Results of the National Achievement of Educational Progress (draft)

## Introduction

The *No Child Left Behind Act of 2001* (NCLB) requires annual testing in reading (or Language Arts) and mathematics developed by each state and sets goals of having all children at the proficiency level or higher by 2013-2014. With the implementation of high-stakes accountability systems, there has been considerable interest by policy makers, educators, and the general public in the confirmation of student progress on state assessments with national benchmarks like NAEP.

Over a decade, numerous published and unpublished studies and reports discuss results, issues, and methodology in comparing test scores on state assessments to NAEP results (Linn and Kiplinger, 1994; Ercikan, 1998; McLaughlin and Banderia, 2003; Linn, 2005; Rothstein, Jacobsen & Wilder, 2006; Braun and Qian, 2007). Perhaps the most noticeable and controversial example is the *Mapping 2005 State Proficiency Standards onto the NAEP Scales*, a recently released research report by the National Center for Educational Statistics (NCES, June, 2007). This report concludes that “although there is an essential ambiguity in any attempt to place state standards on a common scale, the ranking of the NAEP score equivalents to the states’ proficiency standards offers an indicator of the relative stringency of those standards” (iii). In response to the NCES report, the Council of Chief State School Officers (CCSSO) raised serious concerns “about how it may be interpreted through simplification, distortion, or erroneous conclusions..... It is disappointing to see individuals use the results of a narrow study to criticize states for working as hard as they can to comply with a one-size fits all law.” The discussion surrounding the conclusions of the 2005 mapping study perceptibly centers on the validity of such comparisons; how to validly compare state assessment results with those on NAEP; and how to accurately and effectively use NAEP data as external criterion to validate student achievement on statewide assessments.

## Methods of the Study

### 1. Objectives of the Study

The current study explores validity evidence in comparing results from distinct tests. Specifically, the study has three incorporated objectives: (1) discuss validity issues of using NAEP results as a source of confirmatory evidence to validate student achievement on state assessments; (2) identify validity evidence that helps support or refute the validity argument in NAEP/state assessments comparisons; and (3) propose three approaches that may provide more appropriate comparisons.

## 2. Information/Data Collection of the Study

The information/data was collected from published reports, documentations, released data from NAEP and state educational agencies, and empirical research and presentations. Seven states, California, Delaware, Louisiana, Missouri, New Jersey, South Carolina, and West Virginia, were used as examples for analyses in this study based on three criteria:

(1) Accessibility of documentation and reports for the 2005 state assessment program and accountability system for a valid comparison;

(2) Geographic location of the state; and

(3) The states that are used in the study of *Mapping 2005 State Proficiency Standards onto the NAEP Scales* (Braun and Qian, 2007) with varying stringency levels of performance standards.

According to the 2005 mapping study, “for each of the four subject and grade combinations, the NAEP score equivalents to the states’ proficiency standards vary widely, spanning a range of 60 to 80 NAEP score points. Although there is an essential ambiguity in any attempt to place state standards on a common scale, the ranking of the NAEP score equivalent to the states’ proficiency standards offers an indicator of the relative stringency of those standards” (Braun and Qian, 2007, iii). In grade 4 reading, the estimated NAEP score equivalents show that South Carolina is ranked the 2<sup>nd</sup> out of thirty-two states; California the 7<sup>th</sup>, Louisiana the 16<sup>th</sup>, New Jersey the 12<sup>th</sup>, and West Virginia is ranked the 25<sup>th</sup>. In grade 8 reading, the estimated NAEP score equivalents show that South Carolina is ranked the 2<sup>nd</sup> out of thirty-four states; California the fifth, Louisiana the 11<sup>th</sup>, New Jersey the 14<sup>th</sup>, Delaware the 24<sup>th</sup>, and West Virginia is ranked the 30<sup>th</sup>. In grade 4 mathematics, the estimated NAEP score equivalents show that South Carolina is ranked the 4<sup>th</sup> out of thirty-three states; Missouri the 5<sup>th</sup>, Louisiana the 16<sup>th</sup>, New Jersey the 20<sup>th</sup>, and West Virginia is ranked the 27<sup>th</sup>. In grade 8 mathematics, the estimated NAEP score equivalents show that Missouri is ranked the 1<sup>st</sup> out of thirty-six states; South Carolina the 2<sup>nd</sup>, Delaware the 12<sup>th</sup>, New Jersey the 15<sup>th</sup>, Louisiana the 25<sup>th</sup>, and West Virginia is ranked the 33<sup>rd</sup> (Braun and Qian, 2007).

## 3. Organization of the Study

This paper is organized into three main sections. In the first section, validity issues in comparing scores on state assessments with the results of NAEP NAEP/state assessment comparisons are discussed. The elements of validity for NAEP/state assessment comparisons are identified in the second section as well as the guideline for collecting validity evidence. Examples from NAEP and selected states are utilized to support the validity argument for such comparisons. Three approaches for NAEP/state assessment comparisons are described in the third section. Applications of these approaches are demonstrated to foster discussion in the context of validity and efficacy. The three approaches are:

(1) The *Three-Level Content Link Approach* (Zhang et al, 2004, 2007) involves the comparison of state content standards with NAEP frameworks for alignment, the comparison of test specifications between NAEP and state assessments for overlapping test content, and the comparison of sample test questions and scoring rubrics from each test for the similarity of test structure. The content linkage provides necessary validity evidence for the NAEP/state assessment comparison and follow-up statistical procedures (e.g., statistical linking, longitudinal and trend analyses).

(2) The *Statistical Linking on a State-by-State Basis* has been recommended by many researchers (Feuer et al, 1989; Mislevy, 1992; Linn, 2005) to compare student achievement on state assessments with the results of NAEP. Linn (2005) indicates that more analyses can be done on the state-by-state bases given the diversity of state assessments and metrics used for reporting. This type of linkage could provide more accurate and meaningful comparison if the two tests measure the similar construct.

(3) The *NAEP-Like Performance Standards Method* was introduced by Nellhaus in 2000. (Student performance standards on the National Assessment of Educational Progress: Affirmations and Improvements, 2000). In this approach, the similarity of performance standards from various states and NAEP are based on the number and titles of achievement levels, and the method used for standard setting. The probable range of the percentage of students at each achievement level is estimated by using the standard error associated with each test. To improve the validity of NAEP/state assessment comparisons, modifications are proposed in this study.

### Validity of NAEP/State Assessment Comparisons

#### 1. Validity Issues in Comparing Distinct Tests

Validity is the most fundamental consideration in developing and evaluating tests (Standards for Educational and Psychological Testing, 1999). Validity refers to the degree to which theory and cumulative evidence support intended uses of test scores.

It has been a challenge to the educational measurement community to treat scores obtained from distinct tests as if they are interchangeable or, at least, comparable (Linn, 2005). Various linking approaches have been defined and discussed in the literature, such as calibration, projection, and moderation (Mislevy, 1992). Perhaps the most enduring and frequently cited example is the linkage between the two college entrance examinations, ACT and SAT, due to great demands of the public (Dorans et al, 1997; Dorans, 1999; Pommerich et al, 2000; Hanson et al, 2001; Kolen and Brennan, 2004). Another well-known example is the development of linking relationship between NAEP and the Third International Mathematics and Science Study (TIMSS) (Johnson et al, 1998a, 1998b, 2002). Recognizing the importance of overlapping content coverage for linking quality, Johnson et al. (1998) conducted a content validation. They concluded that the sufficiently similarities of test content between NAEP and TIMSS warrant the

linkage as a valid comparison. Mislevy (1992) and Linn (1993) point out that linking test scores from different assessments through statistical procedures must satisfy certain requirements to support interpretable and valid comparisons. The accuracy of this kind of linkage heavily depends on the context of the assessments, the groups used for calculating statistics, and the time of administering the tests (Linn, 1993). More importantly, the two linked tests must measure similar constructs; otherwise, scaling is merely a mathematical operation applied to two sets of data to match test score distributions (Dorans et al, 1997; Dorans, 1999).

The Mapping approach has been used in recent years to transform the proficiency levels for state assessments to the NAEP scales in reading and mathematics and compare the stringency of the state performance standards with the NAEP achievement levels (McLaughlin, 1998; McLaughlin et al, 2003, 2005; Braun and Qian, 2005, 2007). The percentages of proficient students from the NAEP sample schools on statewide assessments are estimated based on the data obtained from a National Longitudinal School-Level State Assessment Score Database (SSASD). McLaughlin and colleagues (2005) declare that “there are many technical reasons for different assessment results from different assessments of the same skill domain. The analyses in this report have been designed to eliminate some of these reasons, by (1) comparing NAEP and state results in terms of the same performance standards, (2) basing the comparisons on scores in the same schools, and (3) removing the effects of NAEP exclusions on trend” (xiv). Kingsbury et al (2005) indicate that the mapping studies investigating the variability in state performance standards with NAEP “suffer from three major limitations” (p. 2): (1) NAEP tests are not designed to align with the content standards of any state; (2) NAEP tests are no-stakes for students and teachers; and (3) the analysis of the mapping studies relies on extrapolation from group performance data for comparisons rather than collecting individual student data from matched group of students. These limitations reduce the strength of the results of comparisons and create “a risk that NAEP results may underestimate the actual level of student performance” (p. 2). Ho and Haertel (2007) point out that the interpretations of those mapping results depend crucially on the often untested assumption that NAEP and state assessments are equivalent. Such comparison of performance standards becomes incoherent and misleading if the tests do not function to measure the same achievement domain. This mapping affords an evaluation of state performance standards as ‘higher’ or ‘lower’ than other states or NAEP that is probably not entirely valid. Reasonable equivalence between state tests and NAEP may or may not exist. The substantial differences between state tests and NAEP will render the mapping illogical and are subject to drift over time.

With the mapping approach, the stringency of state performance standards is compare with the NAEP achievement levels. Do NAEP and state assessments measure the same or similar construct? The authors of the mapping studies acknowledge that their comparisons “are based purely on results of testing and do not compare the content of NAEP and state assessments” (McLaughlin et al, 2005, xi). They admit that “ideally, the quantitative analysis should be supplemented by an intensive examination of the degree of alignment between the state test frameworks and the NAEP frameworks. This has not been done” (Braun and Qian, 2007, p.16). Table 1 lists the states included in the 2003 mapping studies by McLaughlin et al. In cases where the 4<sup>th</sup> or 8<sup>th</sup> grade data was not available from the SSASD, test results from

adjacent grades were used or aggregated test results across elementary grades or middle school grades were used instead. For example, the analyses for 15 states (30%) are conducted by using either grade 3 or grade 5 test results for grade 4 reading (e.g., Delaware, Minnesota); the analyses for 13 states (25%) are based on adjacent grades data for the grade 4 mathematics (e.g., Arizona, Indiana). The assumption behind it is that “because reading [or mathematics] achievement scores for different grades in a school are normally highly correlated with each other, so NAEP grade 4 trends can be compared to state assessment grade 3 or grade 5 trends, and NAEP grade 8 trends can be compared to state assessment grade 7 or grade 9 trends” (McLaughlin et al, 2005, p.7). The authors, however, admitted in their report that more discrepancies between NAEP and state assessment results are to be expected when they are based on adjacent grades, not the same grade, primarily because such comparisons involve different cohorts of students. It is obvious that such comparisons are invalid due to the great variations of content standards, grade level expectations, test construct (knowledge, skills, and cognitive complexity), performance standards, and student population from grade to grade.

To avoid using off-grade data in the 2005 mapping study (Braun and Qian), 18 states (e.g., Arizona, Virginia) are excluded from analyses for grade 4 reading, 17 states (e.g., Alabama, Oregon) for grade 4 mathematics, 16 states for grade 8 reading (e.g., Kentucky, Washington), and 15 states are excluded for grade 8 mathematics (e.g., California, Vermont). The questions remain, however; (1) if the state assessments used for the 2005 mapping study measure the same or similar construct as NAEP as well as with each other; and (2) if there is sufficient validity evidence to support the conclusions of the mapping studies.

## 2. Elements of Validity Evidence in NAEP/State Assessment Comparisons

As stated by Messick (1992), the validation process involves accumulating evidence for and examining potential threats to the validity of test score uses and interpretations. The Standards for Educational and Psychological Testing (1999) outlines the sources of validity evidence that might be used in evaluating a proposed interpretation of test results, such as the evidence based on test content, response processes, internal structure, relations with other variables, and consequences of testing. A sound validity argument should integrate various validity elements into coherent evidence to support intended interpretations of test results for specific uses. What validity evidence is needed for the comparison of student achievement on state assessments with the results of NAEP? How to validly compare state performance standards with the NAEP’s achievement levels? In this study, validity elements for NAEP/state assessment comparisons are classified into four general categories: (1) Purpose of Testing, (2) Test Construct, (3) Test Administration, and (4) Consequences of Testing (Chart 1). For each category, questions and examples are provided as guidelines for collecting validity evidence. It should be noted that this paper does not provide exhaustive coverage of validity evidence, but for the NAEP/state assessment comparisons only.

## (1) Purpose of Testing

“Validity logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use” (Standards, 1999, p. 9). The purpose of testing is a critical issue in validation. Three elements are identified in this study to contribute to the purpose of testing: (a) the primary objectives of the test; (b) the intended uses and interpretations of test results; and (c) the defined population (Chart 1).

The National Assessment of Educational Progress (NAEP) is the only measure of student achievement in the United States where you can compare the performance of students in each state with the performance of students across the nation or in other states. When NAEP is conducted at the state level, results are currently reported separately for public school students only and are broken down by several demographic groupings of students and also reported for the nation. NAEP has two major goals: (a) to compare student achievement in states and other jurisdictions; and (b) to track changes in achievement of 4<sup>th</sup>, 8<sup>th</sup>, and 12<sup>th</sup> graders over time in selected content domains (e.g., reading, mathematics) (The Nation’s Report Card: National Center for Education Statistics, 2007). NAEP is administered biennially to a representative sample of each state.

Under the NCLB requirements state assessments must be administered annually to students of grades 3 to 8 and one grade in high-school in reading (or English Language Arts) and mathematics. Test results are reported at various levels (e.g., district, school, and individual student) and broken down by sub-groups. According to the federal regulations, test results of state assessments are used for high-stakes school accountability. Schools that fail to make the Adequate Yearly Progress (AYP), such as test 95% of students for all sub-groups; and attain the target percent of proficient students, may face the consequences of sanctions. Under the state educational policies, state assessments serve a wide variety of purposes. For example, the primary goal of the California Standardized Assessment Program (STAR) is to help measure how well students are learning required academic skills. The STAR results can be used (a) to provide parents and guardians with information about their children’s progress; (b) as a tool for helping parents, guardians, and teachers work together to improve student’s learning; (c) to help school districts and schools identify strengths and areas that need improvement in their educational program; and (d) to allow the public and policy makers to hold public schools accountable for student learning (California Standards Tests Technical Report – Spring 2005 Administration, 2006). Louisiana Educational Assessment Program (LEAP) is designed to measure how well a student has mastered the state content standards. Louisiana’s high-stakes testing policy is an important part of Reaching for Results, an educational reform system, to improve student achievement. The LEAP tests are developed to ensure that grade 4 and grade 8 students have adequate knowledge and skills before moving on to the next grade. LEAP results are not only used for high-stakes school accountability, but also are tied to promotional policy for individual students, such as grade to grade promotion, intensive summer remediation program (<http://www.louisianaschools.net>; LEAP Assessment Guide, English Language Arts, Mathematics, Science, and Social Studies Grade 4, Louisianan Department of Education;

Revised 2006). As a result of the state Outstanding Schools Act of 1993, the 2005 Missouri Assessment Program (MAP) was designed to identify the knowledge, skills, and competencies that Missouri students should acquire by the time they complete high school and to evaluate student progress toward those academic standards. The Department uses the information obtained through MAP to monitor the progress of Missouri's students in meeting the standards, to inform the public and the state legislature about student's performance, and to help make informed decisions about educational issues (Guide to Interpreting Results, Revised 2005).

The fundamental difference between NAEP and state assessments is the purpose of testing, specifically the intended uses of test results from the target population. NAEP is a national survey of student achievement; while state assessments are designed for high-stakes school accountability under the federal requirements. NAEP results are generated at the national and state levels; while state assessments must report individual student achievement. In addition, the proposed uses and interpretations of test results show great variations among states, according to the state educational policies, from making high-stakes decisions for individual students to providing student achievement information for parents, teachers, policy makers, and the public.

## (2) Test Construct

Validity evidence on the similarity of test construct is essential in the NAEP/state assessment comparison to support the proposed uses and meaningful interpretations of the results of analysis. In this study, the sources of evidence for test construct are delineated in five elements: (a) grade level; (b) test content (e.g., knowledge and skills, cognitive complexity), (c) test structure (e.g., item formats, scoring rubrics and process), technical quality (e.g., reliability, standard error of measurement), and performance standards (e.g., definitions of achievement levels, achievement level descriptors, method and procedure for setting cut scores).

The NAEP Reading Framework sets forth a broad definition of "reading literacy" that includes developing a general understanding of written text, thinking about it, and using various texts for different purposes. In addition, the NAEP framework views reading as an interactive and dynamic process involving the reader, the text, and the context of the reading experience. NAEP Reading uses multiple-choice, short constructed-response, and extended constructed-response questions (Table 2a). The "Contexts for Reading" dimension provides guidance for the types of texts: (a) Reading for Literary experience, (b) Reading for Information, and (c) Reading to Perform a Task (for grade 8 only) (Table 2b). The "Aspects of Reading" dimension provides guidance for the types of comprehension questions: (a) Forming a General Understanding, (b) Developing Interpretation, (c) Making Reader/Text Connections, and (d) Examining Content and Structure (Table 2c).

Chart 2 displays the content domains measured in reading or English language arts for six states, California (CA), Louisiana (LA), Delaware (DE), New Jersey (NJ), West Virginia (WV), and South Carolina (SC) with brief descriptions. According to the 2005 mapping study, as the

reading data is not available for California and Louisiana, the English language arts data is used for comparisons.

- The CA Standardized Testing and Reporting (STAR) English language arts test measures the standards of reading, writing, and language conventions for both grades 4 and 8. In grade 4, for example, reading takes 51% of the total raw score point for (a) Word Analysis and Vocabulary; (b) Reading Comprehension; and (c) Literary Response and Analysis. Writing takes 44% for (a) Writing Strategies; (b) Written and Oral Language Convention; and (c) Writing Application (California Standards Test Blue Prints for Grade 4 and Grade 8 English language arts).
- The LA Educational Assessment Program (LEAP) grade 8 English language arts, for instance, measures Writing (19%); Reading and Responding (55%); Using Information Resources (14%); and Proof Reading (12%) with multiple-choice, constructed-response, and essay questions (Louisianan Educational Assessment Program – 2005 Technical Report).
- Like the NAEP reading test, the Delaware Student Testing Program (DSTP) grade 8 reading focuses on reading comprehension. Students read three types of written text, Informative, Literary, and Technical to respond multiple-choice, short answer, and extended constructed-response questions at three comprehensive levels, Determining Meaning, Interpreting Meaning, and Extending Meaning (Delaware Student Testing Program - 2005 Technical Report).
- In New Jersey (NJ), the Assessment of Knowledge and Skills (NJ ASK) Language Arts Literary for grade 4 measures Reading and Writing, where reading score takes 53% of the total raw score point and writing score takes 47% (New Jersey Assessment of Knowledge and Skills Technical Report – 2005). The NJ Grade Eight Proficiency Assessment (NJ GEPA) also measures Language Arts Literary, in which reading score takes 67% of the total score point and writing score takes 33% (New Jersey Grade Eight Proficiency Assessment Technical Report – 2005).
- The West Virginia (WV) Educational Standards Test (WESTEST) is designed to measure Reading/Language arts, where the reading component takes 68% and 66% out of the total Reading/Language arts score, respectively for grades 4 and 8 (West Virginia Educational Standards Test Technical Report – 2005 Supplement).
- The South Carolina (SC) Palmetto Achievement Challenge Test (PACT) English language arts, using multiple-choice, constructed-response, and extended constructed-response questions, measures Reading, Writing, and Research for both grades 4 and 8. In grade 8, for example, the reading component takes 50% out of the total raw score point, writing score 40%, and the research component takes 10% (Technical Documentation for the 2005 Palmetto Achievement Challenge Tests of English language arts, mathematics, Science and Social Studies).

In mathematics, both state standards and NAEP frameworks are primarily founded on the National Council of Teachers of Mathematics (NCTM) *Curriculum and Evaluation Standards for School Mathematics* (1989), so that considerable overlap in test content is expected. The major differences between NAEP and state assessments in mathematics may involve (a) the grouping of objectives; (b) the title used for each content category; and most importantly (c) the emphasis or weight given to each content domain. Table 3 shows the five content areas that constituted the 2003 NAEP mathematics test, which applied to both grades 4 and 8: Number Sense, Properties, and Operations (40% for grade 4; 20% for grade 8); Measurement (20% for grade 4; 15% for grade 8); Geometry and Spatial Sense (15% for grade 4; 20% for grade 8); Data Analysis, Statistics, and Probability (10% for grade 4; 15% for grade 8); and Algebra and Functions (15% for grade 4; 30% for grade 8). The three mathematical dimensions for NAEP are: Conceptual Understanding; Procedural Knowledge; and Problem Solving. However, the NAEP test specifications do not provide the number of items in each content category.

Chart 3 presents the content domains measured in mathematics for five states, Louisiana (LA), Missouri (MO), New Jersey (NJ), West Virginia (WV), and South Carolina (SC), with a brief description.

- Similar to NAEP, the LA Mathematics Standards are organized into six content categories: Number and Number Relations; Algebra; Measurement; Geometry; Data Analysis, Probability, and Discrete Mathematics; Patterns, Relationships, and Functions, but separate Algebra from Patterns, Relationships, and Functions. The LA Educational Assessment Program (LEAP) gives the greatest weight to Number and Number Relations (33%) for grade 4 as NAEP does; but more weight to Geometry (22%) and to the combination of Algebra; and Patterns, Relations, and Functions (22%) than NAEP does. With more evenly distributed items, the LEAP grade 8 mathematics shows a similar weight as NAEP by content category (Louisianan Educational Assessment Program – 2005 Technical Report).
- During 2001-2005, the MO Assessment Program (MAP) was designed for each subject area one time in each grade cluster of 3-5, 6-8, and 9-11 (Understanding Your Annual Performance Report, 2005-2006). These assessments were developed for grade spans rather than by the end of each grade. The MAP mathematics was administered to students of grades 4, 8, and 10 in 2005 to measure the knowledge and skills, respectively, for grades 3-5, 6-8, and 9-11. The MAP mathematics testes assess six content domains: Number Sense; Geometric, Spatial Sense, and Measurement; Data Analysis, Probability, and Statistics; Patterns and Relationships; Mathematical System and Number Theory; and Discrete Mathematics, for both grade 4 and grade 8. Thirty-seven percent of the test items are given to Number Sense for grade 4 and 29% for grade 8 followed by 21% of the items for Geometric, Spatial Sense, and Measurement in both grades. The mathematics score is a composite score of Terra Nova (61-62% of the total test items) and state-developed items.
- The NJ Assessment of Skills and Knowledge (ASK) for grade 4 and Grade Eight Proficiency Assessment (GEPA) mathematics tests measure four content domains: Number Sense and Numerical Operation; Geometry and Measurement; Patterns and Algebra; and Data Analysis,

Probability, and Discrete Mathematics. Using multiple-choice and constructed-response questions, ASK gives the greatest emphasis to Number Sense (27% out of 33 items) and an equal weight to the other three content domains (New Jersey Assessment of Knowledge and Skills Technical Report – 2005). GEPA gives an equal weight to Geometry and Measurement; and Patterns and Algebra (28% each out of 36 items) and an equal weight to Number Sense; and Data Analysis, Probability, and Discrete Mathematics (22% each) (New Jersey Grade Eight Proficiency Assessment Technical Report – 2005).

- The SC Palmetto Achievement Challenge Test (PACT) grade 4 and grade 8 mathematics tests measure five content categories that are similar to the NAEP framework. Like NAEP, more weight is given to Number and Operation (26%); and Measurement (21%), followed by Algebra and Geometry (18% for each) in grade 4; more weight is given to Algebra (28%) and Number and Operation; and Geometry (24% of each), followed by Data Analysis and Probability (19%) in grade 8 (Technical Documentation for the 2005 Palmetto Achievement Challenge Tests of English language arts, Mathematics, Science and Social Studies).
- The WV Educational Standards Test (WESTEST) mathematics measures the same five content categories as NAEP for both grades. WESTEST gives the greatest weight to Numbers and Operations (38%) like NAEP for grade 4, but followed by Geometry (19%). Similarly, the greatest weight is given to Algebra (29%) as NAEP for grade 8, but followed by Data Analysis and Probability (21%) (West Virginia Educational Standards Test Technical Report – 2005 Supplement).

The brief comparisons of test content and test structure suggest that there are similarities between NAEP and the sample state assessments in reading and mathematics. However, variations are observed from test to test in content domains, weight of each domain, item format, test length, and scoring process. For reading, the NAEP test centers the reading literacy to develop a general understanding of the written text for different purposes at various comprehensive levels. Four of the five state assessments sampled in this study are designed to measure much broader concepts in English language arts, including reading, writing, and language skills (e.g., convention, speaking) with varying proportion and weight for each. In mathematics, the NCTM Standards create the foundation for the mathematics standards and state assessments nationwide. However, the number of items or the maximum score point assigned to each content domain greatly mirror the state curriculum and expectations for students. Although most state assessments are designed to measure on-grade content based on the state standards and grade-level expectations, the 2005 Missouri MAP is a measure of grade clusters; the grade 4 mathematics actually measures the grade span of 3-5 and the grade 8 mathematics measures the grade span of 6-8 prior to the implementation of NCLB requirements. Moreover, NAEP questions are given to students at more than one grade or age level. These questions are referred to as, for example, between grade 4 and grade 8 in Mathematics (NAEP Cross Grade Questions Information, NAEP Questions Tool Help, 2007).

Under the NCLB requirements, states must set challenging performance standards and achieve the goal of having all children at the proficiency level or higher in reading (or Language

arts) and mathematics by 2013-2014. These performance standards indeed play a critical role in state assessment programs and accountability systems. Given the importance of such performance standards as tools for educational policy; the activities of setting cut scores are far more than a methodology or a measurement issue. In addition to meeting the federal requirements, performance standards, such as the number of achievement levels, the definition of each level, and the intended uses, must satisfy the state political agenda. As described by Reckase (2006), standard setting is the process of translating the policy definitions into the most comparable test scores on the reporting scale. This process depends on the understanding of panelists on the state content standards, expectations for students, and the policy definitions. Furthermore, to set challenging, but attainable performance standards, the state educational history, components of student population, geographic location of schools (e.g., urban, suburban, rural), and most importantly the high-stakes consequences of state assessments must be taken into consideration. These cut scores, based on state-developed or state-adopted assessments; reflect the expectations of educators, parents, general public, and policy makers of the state.

As indicated earlier, NAEP is a national survey of student academic achievement in selected subject areas and grade levels. The current NAEP cut scores used to report achievement levels were set in 1990 - 1992 for 4<sup>th</sup> and 8<sup>th</sup> grade reading and mathematics. According to The Status of Achievement Levels of The Nation's Report Card, "the 2001 reauthorization law requires that the achievement levels be used on a trial basis until the Commissioner of Education Statistics determines that the achievement levels are reasonable, valid, and informative to the public.....based on the congressionally mandated evaluation so far, NCES agrees with the National Academy's recommendation that caution needs to be exercised in the use of the current achievement levels. Therefore, NCES concludes that these achievement levels should continue to be used on a trial basis and should continue to be interpreted and used with caution" (National Center for Educational Statistics, Latest updated 03 February 2005; <http://necs.ed.gov/nationasreportcard>).

The mapping studies assume that the 'primary standard', such as 'proficient' or 'meets the standard', is generally the standards that states use for reporting AYP; therefore, they are analogous across states and equivalent to the NAEP's Proficiency Level. Is the assumption valid? Below are examples of the definitions of achievement levels from NAEP and two state assessments.

- NAEP reports three achievement levels: Advanced, Proficient, Basis and reporting includes % of students Below Basic. The definitions of these levels are (2005 NAEP Mathematics Assessment and Item Specifications, National Assessment Governing Board, U.S. Department of Education, 2005, p.7):

Advanced: Superior performance

Proficient: Solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter,

knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Basic: Partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

- Louisiana (LA) uses five achievement levels to report student performance on the LA Educational Assessment Program (LEAP), Advanced, Mastery, Basic, Approaching Basic, and Unsatisfactory (LEAP Assessment Guide, Louisiana Department of Education, revised in September 2006). Below are the general definitions for these levels:

Advanced: A student at this level has demonstrated superior performance beyond the level of mastery.

Mastery: A student at this level has demonstrated competency over challenging subject matter and is well prepared for the next level of schooling.

Basic: A student at this level has demonstrated only the fundamental knowledge and skills needed for the next level of schooling.

Approaching Basic: A student at this level has only partially demonstrated the fundamental knowledge and skills needed for the next level of schooling.

Unsatisfactory: A student at this level has not demonstrated the fundamental knowledge and skills needed for the next level of schooling.

- In South Carolina, the results of the Palmetto Achievement Challenge Tests (PACT) are reported in four achievement levels with the same title as NAEP, Advanced, Proficient, Basic, and Below Basic. (<http://ed.sc.gov/agency/offices/assessment/pact>).

Advanced: The student exceeded expectations for student performance based on the curriculum standards.

Proficient: The student has met expectations for student performance based on the curriculum standards.

Basic: The student has met minimum expectations for student performance based on the curriculum standards.

Below Basic: The student has not met minimum expectations for student performance based on the curriculum standards.

As a national survey, NAEP reading and mathematics measure what students know and are able to do; which are based on the corresponding framework developed for assessments.

There are no consequences of NAEP. The definitions of NAEP achievement levels depict that proficient students perform solid academic achievement of tested grade and demonstrate competency of higher cognitive skills over challenging subject-matter, such as analysis and applications; while students at the Basic level only show partial mastery of prerequisite skills. The state achievement levels, due to the purpose of testing and the intended uses of test results, are closely tied to the state curriculum and grade-level expectations, particularly the readiness of students for the next grade level. These definitions usually differentiate well-prepared students from minimally prepared students for high-stakes decisions, such as promotion and retention. In Louisiana, for example, students at the Mastery Level are well prepared for the next level of schooling; while the Basic Level denotes that students demonstrate fundamental knowledge and skills needed for the next level of schooling. Similarly in South Carolina, students at the Proficient Level meet the expectations based on the state curriculum standards; while students at the Basic Level meet the minimum expectations. The South Carolina “Education Accountability Act requires that schools develop individual Academic Plans for Students (APS) for those students in grades three through eight who score below the Basic level on the PACT or who otherwise lack the skills to perform at grade level” (<http://ed.sc.gov/agency/offices/assessment/pact>).

In the early 1990s, the Committee on Equivalency and Linkage of Educational Tests (1998) pointed out that transforming state test scores to the NAEP achievement levels would produce results with substantial practical ambiguity. Although some states use the same or similar number of achievement levels and/or the same labels as NAEP or with each other, simply assuming that these achievement levels are equivalent or parallel lacks validity evidence. As indicated earlier, setting challenging, but attainable performance standards must consider the state educational history and the high-stakes consequences of testing. For example, the results of a 2007 study to link the 2003 Delaware Student Testing Program (DSTP) to the 2003 NAEP 8<sup>th</sup> grade Mathematics (Zhang, Kersteter, Foret and Wang, 2007) indicate that the three NAEP cut scores are located at the percentile rank of 30, 74, and 96, respectively, for the Basic, Proficient, and Advanced levels and the equipercentile linking functions on the DSTP scale are 473 for Basic, 520 for Proficient, and 572 for Advanced. If the cut score were changed from the current 493 to 520 for the level of *Meets the Standard* for the DSTP mathematics, over 70% of the Delaware 8<sup>th</sup> graders would have failed the test in the sixth year of the DSTP administration. The consequences would be substantial, according to the Delaware high-stakes policy, at the student (e.g., Individual Instruction Plan, remedy programs), school, district (e.g., AYP, accountability), and the state level (e.g., budget).

### (3) Test Administration

Test administration is an important component in educational testing. Standardized procedures ensure the accuracy and comparability of score interpretations and provide equal opportunity for all students (Standards, 1999). However, modified procedures are often needed and accommodations are granted for English Language Learners (ELL) and students with disabilities (SD). Four elements of validity are identified under the category of test

administration. They are testing conditions, accommodation, administration mode, and reporting levels.

“NAEP is a timed assessment administered in English to group of students. Timing is a critical component of standardizing an assessment across the country” (NAEP 2005 AA Manual). The 2005 NAEP reading has ten 25-minute blocks with one or two reading passages accompanied by a set of comprehension questions per block. Each student’s test booklet contains two blocks. The NAEP mathematics consists of ten 25-minute blocks of mathematics questions. Each test booklet contains two blocks (The Nation’s Report Card). The state assessment, however, is generally an untimed test administration. For example, in Louisiana, the Test Administration Manual of LEAP states that “the goal is for each student to complete the test without being constrained by time limits.” However, “all tests and sessions must be administered and completed on the day they are scheduled” (p.9). In West Virginia, the administration of the WESTEST is also untimed. It is suggested that “all sessions of a content area test are to be completed on the same day. Any students who require additional time must be accommodated and allow students to be given the time needed within the confines of the test day” (Examiner’s Manual, p.4).

Same to state assessments, the 2005 NAEP permits the use of calculators for 4<sup>th</sup> and 8<sup>th</sup> grade mathematics; which comprise one-third test items of the test. To align the assessment to the content standards and classroom instruction, many states also provide students with equations and formulas and/or manipulative and tools for mathematics assessments. For instance, mathematics punch-out tools (e.g., ruler for grade 4; ruler and protractor for grade 8) are provided in West Virginia with testing materials and students are asked to be prepared prior to the day the test is administered (Examiner’s Manual, p.19). In Louisiana, “students are not required to recall formulas or unit conversions from memory. A separate Mathematics Reference Sheet containing grade appropriate formulas and equivalencies needed to solve measurement or geometry items is provided. Students are expected to select the proper formula or conversions needed to solve a given problem” (Test Administration Manual, p.2-2).

“The NAEP program has established procedures to include as many students with disabilities (SD) and English Language Learners (ELL) as possible in the assessments. School staff make the decisions about whether to include an SD or ELL student in a NAEP assessment, and which accommodations, if any, they should receive” (NAEP Inclusion Policy). In 2005, the NAEP exclusion rates varied among jurisdictions due to three major factors: (a) the percentages of identified SD and ELL students vary; (b) some required accommodations (e.g., using a calculator for computation, passage translation) would be inconsistent with NAEP’s frameworks; and (c) severe disabilities for some SD students or lacking of English language for some ELL students. Table 4 presents the percentage of identified SD and ELL students combined and the percentage of those students were included in the 2005 NAEP for six states (2005 Nation’s Report Card, National Center for Educational Statistics) and the inclusion rate on the state assessments. For example, in Delaware, the NAEP inclusion rate was 35% for grade 4 and 87% for grade 8 reading; the inclusion rate was 60% for grade 4 and 39% for grade 8 mathematics. In California, the inclusion rate was 89% for reading of both grades; 90% for grade 4 and 89% for

grade 8 mathematics. Under the NCLB requirements, schools must assess at least 95% of eligible students in the state assessment for every sub-group. In Delaware, all students enrolled in a public school are counted as eligible to take the state assessment unless they participate in the alternate assessments (1%) or are granted a special exemption. In 2005, the DSTP participation rate was 98% for SD students and 99% for ELL students in reading; the participation rate was 98% for SD students and 99% for ELL students in mathematics. In California, the state reported that the 2005 participation rate was 99% for ELL students and 98% for SD students in both English language arts and mathematics based on the enrollment of the first day of testing (State Adequate Yearly Progress Report – 2005 Accountability Progress Report).

The usefulness and interpretability of test scores require that directions to test takers, testing conditions, and scoring process follow the same detailed procedures (Standards, 1999). Testing conditions differ widely between NAEP and state assessments in several aspects. NAEP reading and mathematics are administered during a 6-week testing window usually from late January until early March. Each student is assessed in one subject with one or two blocks of the test. Each mathematics test booklet, for example, contains two 25-minute blocks of cognitive testing and one 20-minute background section on content related questions and family background. Thus, NAEP may not furnish enough information to support even minimally valid and reliable scores for individual students. State assessments are mostly administered in spring with a fixed schedule and every student must take the full length of the test except embedded field test items. Each test comprises multiple sessions with one or two sessions per day. The state assessment is generally untimed. Even though reasonable time is recommended for each session, extended time is always permitted upon request. To meet the AYP requirements, each state must ensure that at least 95% of the students with disabilities and English Language Learners participate in the state assessment. With the cooperation of districts and schools, various accommodations are granted to increase the accessibility of state assessments for special populations. To align with the state curriculum and the objectives of state assessments, most states provide supportive materials (e.g., tools, formula sheet) in order to measure the capacity of problem solving. Additionally, the similarity of scoring rubrics for constructed-response questions and the procedures for scoring are important sources of validity evidence in the NAEP/state assessment comparison.

#### (4) Consequences of Testing

Validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Messick points out that “validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other mode of assessment” (1989, p.13). Messick further distinguishes the validity evidence into evidential and consequential basis to support the adequacy and appropriateness of inferences. He also distinguishes between interpretations of test scores and uses of test results (Linn, 2008). For the purpose of the present study, consequences of testing are discussed, particularly the implications of federal and state educational policies on the test design and proposed uses of test results.

“The National Assessment of Educational Progress (NAEP) is the only nationally representative and continuing assessment of what America’s students know and can do in various subject areas” (NAEP Overview, <http://nces.ed.gov/nationsreportcard>). The National Assessment Governing Board (NAGB), an independent bipartisan group, sets policy for NAEP and is responsible for developing the framework and test specifications that serve as the blueprint for the assessment. Under the NCLB Act, “any state that wishes to receive a Title 1 grant must include in the state plan it submits to the Secretary of Education an assurance that beginning in the 2002-2003 school year the state will participate in the biennial state-level National Assessment of Educational Progress (NAEP) in reading and mathematics at grades 4 and 8.” The regulation clearly states that there will be no rewards or sanctions to states, local education agencies, or schools based on the state NAEP results and participation in NAEP is not a substitute for the state’s own assessment of all students in grades 3-8 reading and mathematics (NAEP and No Child Left Behind, national Center for Educational Statistics). NAEP does not provide scores for individual students or schools; instead, it offers state level summary based on a random sample of students from each state.

Federal educational policies, however, have substantial influences on state assessments particularly in the standards movement. The NCLB requires all states to assess reading (or English language arts) and mathematics for students in grade 3 through 8 and one grade in high school annually. State assessments must align to the rigorous content standards and report student achievement with challenging performance standards. Each state has to identify annual or bi-annual measurable objectives that would lead all students to achieve the proficiency level or higher by 2013-2014. Schools that fail to meet the high-stakes AYP requirements (e.g., 95% participation rate, reach the annual or biennial achievement target) have to face the consequences of sanctions. In addition to the federal requirements, the uses of state assessment results may vary greatly from state to state depending on the state regulations and educational policies. For some states, the test scores provide useful information for improving classroom instruction; for some states, the test results are used for accountability at the district and school levels (e.g., reward, sanction), for teacher evaluation, and making high-stakes decisions for individual students (e.g., Individual Instructional Plan, summer school, grade-to-grade promotion, and graduation). Under such circumstances, state assessments must balance the federal’s one-size-fits-all policy and the state regulations, serve multiple purposes, set challenging, but attainable, performance standards to support and promote teaching and learning in order to achieve the ultimate goal of improving student achievement.

- The Louisiana Educational Assessment Program (LEAP) is high-stakes for individual students. The LEAP tests are designed to ensure that grade 4 and grade 8 students have adequate knowledge and skills before moving to the next grade. They set a moving target for student promotion. As of spring 2000, no 4<sup>th</sup> and 8<sup>th</sup> grade student can be promoted if he or she does not achieve *Approaching Basic* or above on both the LEAP English language arts and the LEAP Mathematics tests. As of spring 2004, grade 4 students are required to score *Basic* or above on either the English Language Arts or the Mathematics test and *Approaching Basic* or above on the other to progress to grade 5. As of spring 2005, grade 8 students are required to score *Basic* or above on either the English Language Arts or the mathematics test

and *Approaching Basic* or above on the other to progress to grade 9. Intensive summer remediation must be offered to students who do not score at the achievement level required for promotion, and those students have the opportunity to retest after remediation concludes in the summer” (Section 1: The Louisianan Educational Assessment Program, the Department of Education).

- In Missouri, “from the inception of the MAP [Missouri Achievement Program] through the 2004-2005 school year, the MAP assessments were administered to students for each subject area one time in each grade span (3-5), (6-8), and (9-11).” “These grade span assessments measure student achievement based upon five achievement levels: Step 1, Progressing, Near Proficient, Proficient, and Advanced” (Understanding Your Annual Performance Report (APR), 2006-2007, p.3). It is the primary responsibility of school districts to administer state-required tests and other tests to measure academic achievement and use disaggregated and longitudinal assessment data to adjust its curriculum and instruction. (Integrated Standards and Indicators Manual: Accreditation Standards for Public School Districts in Missouri). The improvement in MAP performance at the district level is analyzed using the past five-year data in the Annual Performance Report. Indicators other than student scores on the state assessment, such as ACT scores and graduation rate; as well as non-academic indicators, such as advanced courses enrollment (e.g., advanced courses, vocational courses, college placement, and vocational placement), student attendance, and dropout rate can be used in the Bonus Point Calculation for district accountability, (Understanding Your Annual Performance Report, 2005-2006 Version 3). In 2005, accreditations were issued to school districts by the state based on the calculation of using academic and non-academic indicators; but no any high-stakes consequences of testing attached to either school or individual student levels in Missouri.
- The New Jersey Grade Eight Proficiency Assessment (GEPA) “serves as a primary indicator for identifying those students who may need instructional intervention in the three content areas of Language Arts Literacy, Mathematics, and Science. The test also serves as an indicator for determining which local education programs may require revisions to ensure that instructional programs are aligned with the Core Curriculum Content Standards. The GEPA is designed to evaluate the progress students are making in mastering the knowledge and skills required by the end of eighth grade. Also, the GEPA provides an indication of students’ progress in the skills required to pass the High School Proficiency Assessment” (2005 Grade Eight Proficiency Assessment Technical Report, March Administration, New Jersey Department of Education; p.3).
- In Delaware, test scores on the Delaware Student Testing Program (DSTP) are used as the primary indicator for high-stakes decisions for individual students; and also for school and district accountability. Student’s whose performance is at Level One *Well below the Standard* for grades 3, 5, and 8 reading and grade 8 mathematics are mandated to attend summer school and re-take the DSTP by the end of summer school unless they met other academic indicators. If the re-test score remains in the lowest level, promotion only can be issued for the student who satisfies other academic indicators to demonstrate the readiness

for the next grade. If the re-test score improves to Level Two *Below the Standard*, the student is promoted with an Individual Improvement Plan (IIP) without further determination; if the test score achieves to the level of *Meet the Standard* or higher, the student is promoted. Students of grade 2 through grade 8, whose reading scores or students of grade 6 through grade 8 whose mathematics scores fall into the level of *Below the Standard* must have an IIP for academic improvement. Students who achieve the highest achievement level of *Distinguished* in any single test are awarded a Distinguished Performance Certificate by the state. The 600 top students in grades 8 and 10 reading, writing or mathematics are eligible for the state Michael C. Ferguson Scholarship. “In 2004 changes were made to the accountability system that would not only meet the NCLB requirements but also incorporated elements of Delaware’s original accountability system” (Educational Accountability – A Partnership of School, Community, and Family, p. 5). “School ratings are given based on their student’s performance on the DSTP,” in seven levels: Superior, Commendable, Academic Review, Academic Progress, Academic Progress – Under Improvement, Academic Watch, and Academic Watch – Under Improvement. Superior or Commendable schools and districts receive a reward; schools or districts do not meet AYP must face the consequences from an Under School Improvement Plan (USI) or Under District Improvement Plan (UDI) to Restructuring Implementation.

- In the General Provisions of South Carolina Education Accountability Act of 1998 “accountability, as defined by this chapter, means acceptance of the responsibility for improving student performance and taking actions to improve classroom practice and school performance by the Governor, the General Assembly, the State Department of Education, colleges and universities, local school boards, administrators, teachers, parents, students, and the community” (Chapter 18, Education Accountability Act of 1998). The regulations identify the objective of the statewide assessment program is to promote student learning and measure student performance; and identify areas in which students need additional support; indicate the academic achievement for schools, districts, and the State; satisfy federal reporting requirements; and provide professional development to educators. “Beginning with the 2005 assessment results, the State Department of Education annually shall convene a team of curriculum experts to analyze the results of the assessments, including performance item by item.” The South Carolina Palmetto Achievement Challenge Tests (PACT) is designed and developed primarily for accountability and serve the state comprehensive approach to improve the curriculum and instruction in public schools. The PACT scores are used to measure academic progress by a comparison of the performance for schools, school districts, and the state from year to year. Individual test scores can be used for high-stakes decisions, such as promotion, retention, and summer school assignment, at the discretion of the school or school district administration. (<http://ed.sc.gov/agency/offices/assessment/pact>).

The NCLB Act of 2001 requires states participating in NAEP every other year for reading and mathematics in grade 4 and grade 8 and giving permission for releasing state test results and data. As a national survey of student achievement, the NAEP tests are designed to generate test results at the nation and state levels only. According to the federal policy, the results of NAEP are not to be used for rewards or sanctions to states and local agencies or used

for school accountability. The NCLB clearly states that “the purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments” (No Child Left Behind Act of 2001). More specifically, the NCLB calls for closing achievement gaps, and improving and strengthening accountability, teaching, and learning by using state assessment systems. Under the federal requirements, the results of grades 4 and 8 reading and mathematics on state assessments must be used primarily for accountability. Incorporated the state educational policy with the federal regulations, many states design their assessment programs to serve multiple purposes, from providing information for program evaluation and adjustment of curriculum and instruction, such as in Missouri and South Carolina to using test results for high-stakes decisions for individual students, such as promotion, retention, and mandated summer school in Delaware and Louisiana.

In addition, there are substantial challenges to the use of the (SSASD) database to make valid inferences about program effectiveness. Causal inference is always difficult to justify in the absence of random assignment (Linn, 1995). The variations of state by state assessment data, even year-to-year assessment data within a state are responsible for the variations include modifications of the assessment program, adjustment of cut scores, mobility of student population, and the rules to validate test scores.

### Approaches to Comparing State Assessment with NAEP

To explore the new role of NAEP under NCLB, the National Assessment Governing Board (NAGB) established an Ad Hoc Committee in 2001. Recognizing the diverse landscape of existing statewide assessments in design and difficulty level; and identifying measurement issues and technical challenges, the committee report concludes that, through a careful review of NAEP’s capacity and test results of eight states, NAEP can serve as a source of confirmatory evidence effectively for state test results. The report emphasizes the importance of overlapping content coverage to link state assessments to NAEP and recommends the use of a single state data and common group design to improve the linking accuracy based on empirical evidence from previous studies (Linn and Kiplinger, 1994; Ercikan, 1998; Johnson et al, 1998a, 1998b, 2002).

In this section, three approaches for NAEP/state assessment comparisons are described and discussed in the context of validity and efficacy: the Three-Level Content Link Model (Zhang et al, 2004, 2007, 2008); the Statistical Linking on a State-by-State Basis (Feuer et al, 1989; Mislavy, 1992; Linn, 2005); and the NAEP-Like Performance Standards Approach (Nellhaus, 2000) with examples to facilitate the discussion.

#### 1. Three-Level Content Link Model

The *Three-Level Content Link Model* is the process of collecting information at three levels to determine the similarity of test construct in comparing distinct tests. The information

provides content validity evidence to support the follow-up statistical linkage or other mathematical procedures and combining with other information, to generate meaningful interpretations of the comparison results. The three-level linkage involves:

### (1) Link the Content Standards

At the standards-level linkage, the state curriculum standards (or grade-level expectations) are compared with the NAEP framework for alignment. It is important to note that (a) state content standards are developed for curriculum and classroom instruction; while the NAEP framework are developed for assessment; and (b) the structure and organization of content standards may vary from state to state. The comparison should focus on what knowledge and skills that students are expected to know and able to do by the end of each grade level. A general review by 2-3 content specialists is recommended. The degree of alignment between the NAEP framework and the state content standards could also be evaluated by using the Webb's criteria (1997): (a) *Categorical Concurrence*, (b) *Depth of Knowledge Consistency*, and (c) *Range of Knowledge Correspondence*. However, modifications are necessary in order to meet the needs for the content linkage since Webb's procedure is designed to align the assessment to the standards and all calculations are based on the number or the percentage of items.

### (2) Link Test Specifications

Linking the test specifications between NAEP and state assessments focuses on the similarity of test construct, particularly the overlapping content and comparable difficulty level. The comparison should include the grouping of objectives, content domain categories and cognitive complexity, weight or emphasis of each content domain at various cognitive levels of the test (e.g., the percentage of items), item types, test length, maximum score point, and general scoring procedure.

### (3) Link Test Items

At the item-level linkage, a sample of items is selected from each test. For the best results, sample items should be matched as measuring the same or similar content standards or category, in the same item format, and for the same grade level. The comparison focuses on the similarity of item context, difficulty level or cognitive demand, scoring rubrics for constructed-response items, and scoring process. In some cases, a review of sample responses may help understand the scoring process.

A study of linking the Delaware Student Testing Program (DSTP) to the 2003 NAEP 4<sup>th</sup> and 8<sup>th</sup> grade mathematics (Zhang, Kersteter, Foret, and Wang, 2007) is an example of the application of the Three-Level Content Link Model. The results suggest that one major difference between the NAEP Framework and the Delaware Content Standards in Mathematics is the grouping of objectives. For instance, the Delaware standards separate Estimation, Measurement, and Computation from Number Sense; while NAEP separates Number Sense, Properties, and Operations from Measurement. When combining these two categories, the

proportion of test items contributes to the combined category is nearly the same from the DSTP (53% for grade 4 and 41% for grade 8) and NAEP (60% for grade 4 and 40% for grade 8). Similarly, the Delaware standards separate Algebra from Patterns, Relationships, and Functions; while NAEP's framework groups the two standards into one. When combining these two standards as NAEP, the percentages of items are nearly equal from the DSTP (16%) and NAEP (15%) in grade 4; the percentages are similar between the two tests (31% for DSTP; 25% for NAEP) in grade 8. Both DSTP and NAEP classify the cognitive complexity of test items in three categories with the same labels and the same descriptions. They are Conceptual Knowledge, Procedural Knowledge, and Problem Solving. However, NAEP did not specify the percentage of items in the three mathematical abilities in practice. Both tests use multiple-choice, short answer (0-2 scale) and extended constructed-response questions (0-4 scale) to measure mathematical concept and skills. The item-level linkage provides additional information about the similarities of item context and general scoring rubrics for open-ended items. The discrepancy of performance expectations between the DSTP and NAEP may be due to the purpose of testing and proposed uses of test results. It is found that the majority of the DSTP test items measure on-grade mathematical knowledge and skills for the 4<sup>th</sup> and 8<sup>th</sup> graders; while some NAEP items are given to students at more than one grade or age level. These questions are referred to as, for example, between grade 4 and grade 8 (NAEP Cross Grade Questions Information, NAEP Questions Tool Help, 2007).

## 2. Statistical Linking on a State-by-State Basis

The statistical linkage has been recommended by researchers to compare student achievement on state assessments with the results of NAEP. Linn (2005) suggests that more analyses can be done on the state-by-state bases given the diversity of state assessments and metrics used for reporting. Various types of linkages, such as calibration, projection, and moderation, have been defined and discussed in the literature (Mislevy, 1992; Linn, 1993). Kolen and Brennan (2004) identify four advantages of the equipercentile method over the mean, linear, and parallel-linear methods for linking: (1) linking equivalents are within the range of observed scores to avoid the out-of-range problem; (2) the relationships between linked tests are not assumed to be linear; (3) the cumulative distribution function of *X*-scores is approximated by the cumulative distribution function of *Y*-scores; and (4) the moments for transformed scores are approximately the same. It is important to note that the statistical linking only can be performed in a result of meaningful interpretations when the tests measure the similar construct with overlapping content domains.

As indicated earlier in this paper, the most enduring and frequently cited example is the linkage between the two college entrance examinations, ACT and SAT (Dorans et al, 1997; Dorans, 1999; Pommerich et al, 2000; Hanson et al, 2001; Kolen and Brennan, 2004). The statistical linking is conducted on the basis that the two tests serve the same purpose and measure the similar construct. Using the statistical moderation procedure, Phillips (2007) links the state-by-state results on the 2005 and 2007 NAEP to the 2003 Trends in International Mathematics and Science Study (TIMSS). The author indicates that TIMSS was purposely designed to be linkable to NAEP and both tests "are conducted in the same grades, cover similar content

standards, use matrix sampling of cognitive items, ....., use similar scaling models (item response theory), and use similar analysis models (plausible values) (p.12). The results of linking the Delaware Student Testing Program (DSTP) to the 2003 NAEP 4<sup>th</sup> and 8<sup>th</sup> grade Mathematics (Zhang, et al, 2007) suggest that (a) the two assessments measure similar construct based on the results of the three-level content link; (b) “the property of population invariance is reasonably attained for male and female students” (p.12); and (c) using the concordance tables to estimate Delaware student performance on the 2005 NAEP mathematics, the difference between estimated and the actual average score and percent of proficient students are relatively small.

### 3. NAEP-Like Performance Standards Approach

This approach was first introduced by Nellhaus in 2000 for a research project of the National Assessment of Governing Board (NAGB) (Student performance standards on the National Assessment of Educational Progress: Affirmations and Improvements, 2000). State performance standards are classified into three categories: same as, similar to or different from NAEP achievement level categories based on the number, the title, and very brief descriptions of these levels (e.g., Advanced means superior performance; Proficient means solid performance), and the method used for standard setting. The other criterion used is whether the state reporting of assessment results consistent with NAEP. The assumption of this approach is that if a state assessment program is similar to NAEP, the results reported by the two assessments should be similar. The probable range in the percentage of students at each achievement level is determined by using the standard error associated with each assessment program. Nellhaus (2000) stated that “NAEP typically reports a standard of error of 1 percentage point at the *Advanced* level, and a standard error of 2 percentage points at the *Proficient*, *Basic*, and *Below Basic* levels” (p.109). It is assumed that the standard error for the state assessments is the same as NAEP. Two standard errors were used to recognize the differences in test content, standard setting procedures, and other factors.

The oversimplified procedure seems to ignore the importance of similar construct and overlapping test content when comparing distinct tests. Since the approach was proposed before the implementation of the NCLB Act of 2001, important factors, such as the purpose and consequences of testing, in high-stakes accountability were not under consideration. Modification of the NAEP-Like Performance Standard Approach is recommended in this study to improve the validity of such comparisons.

(1) A general review of the performance (or achievement) level descriptors (PLD) for compared tests should be conducted by a panel of content and assessment specialists. The review process provides the information on what students are expected to know and are able to do at each achievement level, which can be used to determine the likeness between the tests depending on the similarity of expectations from students. The intended uses of these achievement levels, particularly in the high-stakes accountability (e.g., promotion, retention) may provide additional information to identify the comparable achievement levels across tests. Nellhaus’s categories can be adopted to label the degree of similarity for NAEP/state assessment comparisons (e.g., Same as NAEP, Similar to NAEP or Different from NAEP). It is important to

note that the likeness of performance standards between compared tests is based on the expectations from students through professional judgment rather than based on the label of achievement levels.

(2) The calculation of the standard error should be conducted for each test. The result is an interval estimate of the percentage of students reported associated with each achievement level with varying probability of covering the true population values. The application of the standard error should be depending upon the similarity of the performance standards between compared tests. If the state's performance standards are considered *Different from NAEP*, the use of standard error is not recommended for the NAEP/state assessment comparison.

(3) The NAEP-Like Performance Standards approach is a simple procedure that offers a general estimate of the range of percentage students at achievement levels with a given confidence interval. The comparison of the PLD provides the information about the expectations from students by each achievement level; however, the diversity of test construct, administration conditions, and intended uses of test results between NAEP and state assessments or with each other may have immeasurable impact on student performance. Cautions should be considered in the interpretations of the results.

### **Summary and Conclusions**

There has been considerable interest by parents, educators, policy makers, and the general public in the confirmation of student progress on state assessments with NAEP's results, particularly under the NCLB requirements for high-stakes accountability. The current study is to explore validity evidence in comparing student performance on the state assessment with test results of NAEP in general and to investigate validity issues in comparing performance standards of state assessments with NAEP achievement levels in particular.

The protocol of validity evidence that needs to be collected for the NAEP/state assessment comparison is developed in this study in four general categories: Purpose of Testing, Test Construct, Test Administration, and Consequences of Testing. The elements of validity with sample questions presented in Chart 1 outline various sources of evidence that might be used in evaluating the proposed interpretations of the results from NAEP/state assessment comparisons. Seven state assessments used as examples in the discussion not only provide substantial information to confirm the importance of validity in such comparisons, but also identify the diversity between NAEP and state assessments as well as with each other.

The fundamental difference between NAEP and state assessments is the purpose of testing. NAEP is responsible to the National Assessment Governing Board (NGAB) to continue its role as a national survey of student achievement. State assessments, however, must serve multiple objectives to meet the federal requirements and satisfy the state educational policy. The design of assessment programs is determined by the purpose of testing and proposed uses of test results, such as test content, test length, and reporting levels. For example, NAEP uses matrix

item sampling to cover wide content domains and report test results at the nation and state levels. State assessments are closely aligned to the state content standards and curriculum to measure student progress toward the standards and their readiness for the next grade level; as well as hold schools accountable. In the high-stakes testing environment, performance standards play a critical role to determine student achievement and serve as an important tool for educational policy. These performance standards, particularly the cut scores mirror the political definitions, support the intended uses of test results, and most importantly, signify the expectations for students from parents, educators, and the public.

The ultimate goal of the standards-based educational reform is to improve student academic achievement. Does it help accelerate student achievement if all states and jurisdictions set performance standards in the same stringency or as stringent as NAEP achievement levels? Are there any relationships between the stringency of performance standards and student achievement? Table 5 shows the ranking of relative stringency of cut scores based on the 2005 mapping study (Braun and Qian, 2007) for the seven states sampled in this study and the percent of proficient students on the 2007 NAEP, the increase in scale scores and the reduction of poverty gaps from 2003 to 2007 by the recently released Quality Counts 2008 for these states. There is no clear relationship observed between the relative stringency of state performance standards and their performance and achievement on NAEP. The 2005 mapping study also reports a weak relationship between states' NAEP means and their NAEP score equivalents. The correlations between where states set their proficiency standards and how they perform on NAEP are .27 (with a standard error of .176) for grade 4 and .01 (with a standard error of .177) for grade 8 in reading; the correlations are .11 (with a standard error of .179) for grade 4 and .23 (with a standard error of .167) for grade 8 in mathematics.

Validity is the most fundamental consideration in developing and evaluating tests (Standards for Educational and Psychological Testing, 1999). Simply applying a mathematical procedure to transfer state performance standards into the NAEP scale obviously lack validity evidence to support the conclusions and perhaps, create numerous confusions to the public. Although the NCLB regulation clearly states that there will be no rewards or sanctions to states, local education agencies or schools based on the state NAEP results; and the NAEP achievement levels continue to be used on a trial basis and should continue to be interpreted and used with cautions, there has no shortage of studies particularly focus on the stringency of cut scores and large discrepancies of percentage of students at achievement levels. Many measurement experts point out that comparing percentage of students above a cut score is not the best approach and can create misleading findings (Holland, 2002; Linn, 2005, 2008). Moreover, the validity of evaluating the stringency of cut scores that one state is higher or lower than the other states is seldom addressed in these mapping studies, but it should be. And it is particularly true that such comparisons depend on the current status rather than on improvement, even the improvement on NAEP of each state. With the increasingly interest of using NAEP results as a source of confirmatory evidence to validate student achievement on state assessments, it is indeed necessary to develop valid and comprehensive approaches for the NAEP/assessment comparison.

## References

- Braun, H. & Qian, J. (2005). Mapping state performance standards onto the NAEP scale. Paper presented at the ETS Conference of Linking and Aligning Scores and Scales: A Conference in Honor of Ledyard R. Tucker's Approach to Theory and Practice, Princeton, NJ: Educational Testing Service, June 2005.
- Council of Chief State School Officer (2007). CCSSO responds to NCES report. Press Release.
- Dorans, N.J., Lyu, C.F., Pommerich, M., & Houston, W.M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University*, 73(2), 24-34.
- Dorans, N.J. (1999). Correspondences between ACT and SAT I Scores. College Board Report No. 99-1; ETS RR No. 99-2.
- Ercikan, K (1998). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across states. *Applied Measurement in Education*, 10, 145-160.
- Feuer, M.J., Holland, P.W., Bertenthal, M.W., CadelleHemphill, F., & Green, B.F. (1998). Interim Report. Committee on Equivalency and Linkage of Educational Test. National Academy Press, Washington, D.C.
- Hanson, B.A., Harris, D.J., Pommerich, M., Scoring, J.A., and Qing, Y. (2001). Suggestions for the evaluation and use of concordance results. ACT Research Report Series, 2001-1.
- Ho, A. and Haertel, E. (2007). Apple to apple? The underlying assumptions of state-NAEP comparisons.
- Ho, A. and Haertel, E. (2007). Over-interpreting mappings of state performance standards onto the NAEP scale.
- Johnson, E.G., and Owen, E. (1998a). Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A technical report (Publication No. NCES 98-499). Washington, DC: National Center for Education Statistics.
- Johnson, E.G., Siegendorf, A., and Phillips, G.W. (1998b). Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study: Eight grade results (Publication No. NCES 98-500). Washington, DC: National Center for Education Statistics.

- Johnson, E.G. (2002). Linking NAEP 2000 to TIMSS 1999. Paper presented at the 2002 AERA/NCME Annual Conference, New Orleans, LA.
- Kolen, M.J. and Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and Practices* (2<sup>nd</sup> Edition). Springer.
- Lane, S. (1999). Validity evidence for assessment. Paper presented at the 1999 Edward F. Reidy Interactive Lecture Series sponsored by the National Center for the Improvement of Educational Assessment, Inc. Providence, RI, October, 14, 1999.
- Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83-102.
- Linn, R.L. and Kiplinger, V.L. (1994). Linking statewide tests to the National Assessment of Educational Progress: Stability of results (CSE Technical Report 375). National Center for Research on Evaluation, Standards, and Student Testing.
- Linn, R.L. (2005). Adjust for differences in tests. Paper prepared for a Symposium on the Use of School-Level Data for Evaluating Federal Education Programs. Washington, D.C: The Board on Testing and Assessment, The National Academies, December 9, 2005.
- Messick, S. (1992). Validity. In R.L. (Ed.), *Educational Measurement* (3<sup>rd</sup> ed.). New York: American Council on Education.
- McLaughlin, D. (1998). Study of linkages of 1996 BAEP and state mathematics assessments in four states. Washington, D.C.: National Center for Education Statistics.
- McLaughlin, D. and Bandeira de Mello, V. Comparing state reading and math performance standards using NAEP. Paper presented at the CCSSO National Conference on Large-scale Assessment. San Antonio, June, 2003.
- McLaughlin, D., Bandeira de Mello, V., Blankenship, C., Chaney, K., Esra, P., Hikawa, H., Rojas, D., William, P. & Wolman, M. (2005) Comparison between NAEP and state mathematics assessment results: 2003. Final Report, Volume I. American Institute for Research, February, 2005.
- McLaughlin, D., Bandeira de Mello, V., Blankenship, C., Chaney, K., Esra, P., Hikawa, H., Rojas, D., William, P. & Wolman, M. (2005) Comparison between NAEP and state mathematics assessment results: 2003. Final Report, Volume II: Appendix D State Profiles. American Institute for Research, February, 2005.
- Mislevy, R.J. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: Policy Information Center, Educational Testing Service.

National Center for Education Statistics (2007). Mapping 2005 state proficiency standards onto the NAEP scales. Research and Development Report.

National Assessment Governing Board (2000). Student Performance Standards on the National Assessment of Educational Progress: Affirmations and Improvements, NAGB, November 2000.

Nellhaus, J.M. (2000) State with NAEP-Like Performance Standards at the *Student performance standards on the National Assessment of Educational Progress: Affirmations and Improvements*. Washington, D.C: National Assessment Governing Board. November, 2000.

New Jersey Department of Education (2005). Grades 3 and 4 New Jersey Assessment of Skills and Knowledge Technical Report.

Pommerich, M., Hanson, B.A., Harris, D.J., and Sconing, J.A. (2000). Issues in creating and reporting concordance results based on equipercentile methods. ACT Research Report Series, 2000-1.

Rothstein, R., Jacobsen, R. & Welder, T. (2006). 'Proficiency for all' – An oxymoron. Paper prepared for the Symposium, "Examining American's Commitment to Closing Achievement Gaps: NCLB and Its Alternatives," sponsored by the Campaign for educational equity, Teacher College, Columbia University, November, 13-14, 2006.

*Standards for Educational and Psychological Testing* (1999). American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Using the National Assessment of Educational Progress to confirm state test results. A Report of the Ad Hoc Committee on Confirming Test Results, Attachment B. March 1, 2002.

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education (NISE Research Monograph No. 6). Madison: University of Wisconsin–Madison, National Institute for Science Education. Washington, DC: Council of Chief State School Officers.

Zhang, L.R. and Lau, Allen (2003). Linking a statewide assessment to the 1999 TIMSS for eighth grade mathematics. Paper presented at the 2003 AERA/NCME Annual Conference, Chicago, IL, April 21-25, 2003.

Zhang, L.R, Kersteter, P., Foret, K., and Wang, S.D. (2007). Linking a statewide assessment to the 2003 National Assessment of Educational Progress (NAEP) for 4<sup>th</sup> and 8<sup>th</sup> grade mathematics. Paper presented at the 2007 AERA/NCME Annual Conference, Chicago, IL, April 10-12, 2007.

**Table 1. List of States Used in the 2003 Mapping Study**

State	Different Score Used	Adjacent Grade Used for Comparison			
		Reading GR 4	Reading GR 8	Math GR 4	Math GR 8
Alabama	Percentile Rank				
Arizona		5		5	
California			7		7
Colorado				5	
Delaware		5		5	
Hawaii		5		5	
Illinois		5		5	
Indiana		3		3	
Kansas		5			7
Kentucky			7	5	
Maryland		5		5	
Massachusetts			7		
Michigan			7		
Minnesota		3		5	
Missouri		3	7		
Nevada			7		7
New Hampshire		3			
New Mexico	Percentile Rank				
Oklahoma		5		5	
Oregon		5		5	
Pennsylvania		5		5	
Tennessee	Percentile Rank				
Utah	Percentile Rank	5		5	
Virginia		5			5
Washington			7		7
West Virginia	Percentile Rank	E	M		
Total	5	16	8	13	5

In West Virginia, E and M represent aggregates across elementary and middle grades, respectively' (Source: Comparison between NAEP and State Reading Assessment results: 2003; Final Report, Volume 1. McLaughlin et al, AIR, February, 2005. Table 4, p.33)

**Chart 1. Elements of Validity Evidence for NAEP/State Assessments Comparison**

Category	Validity Elements	Questions and Examples
<b>Purposes of Testing</b>	1. Objectives	<p><i>What are the primary purposes of the assessment?</i></p> <p>Measure student yearly progress toward the content standards            Improve classroom instruction            Used as high-stakes accountability test at various levels</p>
	2. Intended uses of test scores	<p><i>What are the intended uses of test scores?</i></p> <p>Provide feedback for teaching and learning            Provide student strengths and weaknesses            Serve as primary indicators for high-stakes accountability            Used for determining the Adequate Yearly Progress            Used for state, district, and school-levels accountability (e.g., reward, sanction, reconstruction)            Used for high-stakes students accountability (e.g., promotion, retention, graduation, and summer school)            Used for teacher evaluation (e.g., reward, sanction)</p>
<b>Test Construct</b>	1. Grade level	<p><i>What is the grade level tested?</i></p> <p>The test is designed to measure by the end of grade or by the end of grade cluster            The test is designed to measure on grade or end-of-course</p>

Category	Validity Elements	Questions and Examples
<b>Test Construct</b>	2. Test content	<p><i>What does the assessment measure?</i></p> <p>Content domain measured (e.g., reading comprehension or English language arts or communication arts)</p> <p>Cognitive complexity measured</p> <p>Weight of each content domain in the test (e.g., percentage of each content domain or standard)</p>
	3. Test structure	<p><i>How is the assessment structured?</i></p> <p>Item type used (e.g., multiple-choice, constructed-response, essay)</p> <p>Test length and maximum score point</p> <p>Scoring rubrics and scoring process for constructed-response items</p> <p>Test form (e.g., single form, multiple forms, matrix sampling)</p> <p>Weighted scoring (e.g., a composite score of reading and writing)</p> <p>Reporting scale used (e.g., developmental scale across grades)</p>
	4. Technical quality	<p><i>What is the technical quality of the assessment?</i></p> <p>Align to the content standards</p> <p>Reliability and standard error of measurement</p> <p>Fairness of testing (e.g., opportunity to learn)</p>
	5. Performance standards	<p><i>What are the expectations for students at the various achievement levels of the assessment?</i></p> <p>Definitions of achievement levels</p> <p>Achievement level descriptors</p> <p>Method and procedure for setting cut scores</p> <p>Accuracy and consistency of classifications</p>

Category	Validity Elements	Questions and Examples
<b>Test Administration</b>	1. Testing conditions	<p><i>Under what conditions the assessment is administered?</i></p> <p>Testing date (e.g., month and year of testing)</p> <p>Testing time (e.g., timed vs. untimed)</p> <p>Test setting (e.g., single session, multiple sessions, multiple days)</p> <p>References and manipulative tools allowed (e.g., dictionary, thesaurus, calculator)</p>
	2. Accommodations	<p><i>How much percent of the students with disabilities and English language learners are included in the assessment?</i></p> <p>Inclusion rules and guidelines</p> <p>The IEP team makes decisions on accommodations) for individual students</p> <p>Available accommodations for students with disabilities and English language learners</p> <p>Aggregation rules applied for reporting</p>
	3. Administration mode	<p><i>How is the test delivered?</i></p> <p>Test delivery device (e.g., paper/pencil, online testing)</p> <p>Standardized testing or computer adaptive testing (CAT)</p>
	4. Reporting test results	<p><i>How the test results are reported?</i></p> <p><i>What are the reporting levels?</i></p> <p>Reporting levels (e.g., state, district, school, class, and student)</p> <p>Reporting by sub-groups (e.g., SPED, ELL, low-income)</p> <p>Reporting participation rate by sub-groups</p> <p>Reporting school yearly progress</p>

Category	Validity Elements	Questions and Examples
<b>Consequences of Testing</b>	1. Federal regulations	<p><i>What are the implications of the federal regulations on test design and intended uses of test results?</i></p> <p>Implications of NCLB requirements (e.g., all students be proficient in school year of 2003-2004 in reading and mathematics, AYP)</p> <p>Peer Review process</p> <p>Provide alternate assessment for students with disabilities</p> <p>Provide alternate assessment for English language learners</p>
	2. State educational policies	<p><i>What are the implications of the state educational policies on test design and intended uses of test results?</i></p> <p>State regulations on assessment (e.g., graduation requirements)</p> <p>State high-stakes accountability system for schools and school districts (e.g., rewards and sanctions)</p> <p>State high-stakes accountability system for individual students (e.g., promotion, graduation)</p> <p>State high-stakes accountability system for teachers (e.g., evaluations)</p>
	3. Student motivation	<p><i>What is student motivation level for the assessment?</i></p> <p>Student participation rate</p> <p>Students are highly motivated to take the assessment</p> <p>Schools motivate students for the assessment</p>

**Chart 2. Item Distributions by Content Domains in Reading/English Language Arts by State**

State	Assessment Program	Grade	Content Domain	Test Item		
				Format	N. (max. pt.) % (pt.) <sup>2</sup>	
California	Standardized Testing and Reporting  (STAR) <i>English language arts</i>	4	1. Reading	MC Task	55 (51)	
			• Word analysis and vocabulary		18 (18)	24 (22)
			• Reading comprehension		15 (15)	20 (18)
			• Literary response and analysis		9 (9)	12 (11)
			2. Writing			45 (49)
			• Writing strategies		15 (15)	20 (18)
			• Written and oral language convention		18 (18)	24 (22)
		• Writing application <sup>1</sup>	1 (8)	(10)		
		Total	76 (83)			
		8	1. Reading		56	
			• Word analysis and vocabulary	9 (9)	12	
			• Reading comprehension	18 (18)	24	
			• Literary response and analysis	15 (15)	20	
			2. Writing		44	
• Writing strategies	17 (17)		23			
• Written and oral language convention	16 (16)		21			
Total	75 (75)					

State	Assessment Program	Grade	Content Domain	Format	Test Item N. (max. pt.)	%
Delaware	Delaware Student Testing Program (DSTP) <i>Reading</i>	8	1. Reading by passage types	MC		
			• Informative	SA	24	38
			• Literary	ECR	24	38
			• Technical		16	25
			2. Reading items by stances			
			• Determining meaning		14	22
			• Interpreting meaning		29	45
			• Extending meaning		21	33
	Total			64 (84)		

State	Assessment Program	Grade	Content Domain	Format	Test Item N. (max. pt.)	% (pt.)
Louisiana	LA Educational Assessment Program (LEAP) <i>English language arts</i>	4	1. Writing (in response to prompt) • descriptive or narrative	Essay	1 (12)	(19)
			2. Reading and Responding • 4 reading passages: 2 short/2 long	MC	20 (20)	(55)
			3. Using Information Resources • One resource packet with 4 to 6 sources	SA	8 (16)	
				MC	5 (5)	(14)
			4. Proofreading • Editing one short passage	SA	2 (4)	
				MC	8 (8)	(12)
			Total		44 (65)	
		8	1. Writing (in response to prompt) • narrative or expository	Essay	1 (12)	(17)
			2. Reading and Responding • 4 reading passages: 2 short/2 long	MC	20 (20)	(58)
				SA	8 (16)	
				ECR	1 (4)	
			3. Using Information Resources • One resource packet with 4 to 6 sources	MC	5 (5)	(13)
			4. Proofreading • Editing one short passage	SA	2 (4)	
				MC	8 (8)	(12)
Total		45 (69)				

State	Assessment Program	Grade	Content Domain	Format	Test Item N. (max. pt.)	% (pt)	
New Jersey	NJ Assessment of Knowledge and Skills (NJ ASK) <i>Language Arts Literacy</i>	4	1. Writing			(47)	
			• write about pictures	Essay	1 (10)	(23)	
			• write about poems	Essay	1 (10)	(23)	
			2. Reading			(53)	
			• reading with text	MC	7 (7)	(16)	
			• analyzing text	MC	4 (4)	(37)	
				OE	3 (12)		
		Total			16 (43)		
	NJ Grade Eight Proficiency Assessment (NJ GEPA) <i>Language Arts Literacy</i>	8	1. Writing				(33)
			• write persuade task to a prompt	Essay	1 (6)	(11)	
			• write speculate task to a picture		1 (12)	(22)	
			2. Reading			(67)	
			• interpreting text	MC	12 (12)	(37)	
				OE	2 (8)		
• analyzing/critiquing text			MC	8 (8)	(30)		
	OE	2 (8)					
	Total			26 (54)			

State	Assessment Program	Grade	Content Domain	Format	Test Item N. (max pt.)	% (pt.)	
West Virginia	WV Educational Standards Test (WESTEST) Reading/Language arts	4	1. Reading • vocabulary and reading comprehension skills for gaining information • performing tasks and reading literacy experience	MC	50 (54)	67 (68)	
				CR			
				2. Writing • language mechanics • language expression		25 (26)	33 (32)
					Total	75 (80)	
		8	1. Reading • vocabulary and reading comprehension skills for gaining information • performing tasks and reading literacy experience	MC	49 (53)	65 (66)	
				CR			
				2. Writing (same as grade 4) • language mechanics • language expression		26 (27)	35 (34)
					Total	75 (80)	

State	Assessment Program	Grade	Content Domain	Format	Test Item N. (max pt.)	% (pt.)
South Carolina	Palmetto Achievement Challenge Tests (PACT) <i>English language arts</i>	4	1. Reading	MC	24 (24)	(56)
				CR	3 (6)	
			2. Writing	MC	4 (4)	(35)
				ER	1 (15)	
				MC	5 (5)	(9)
		Total		37 (54)		
		8	1. Reading	MC	36 (36)	(50)
				CR	5 (10)	
			2. Writing	MC	7 (7)	(40)
				ER	2 (30)	
MC	5 (5)			(10)		
Total	CR	2 (4)				
Total		57 (92)				

<sup>1</sup> The grade 4 Writing Application is not included in the percentage calculation on the web page of California Department of Education.

<sup>2</sup> The percentage is calculated including the writing task for grade 4 in this study.

**Table 2a. Distribution of Items for 2005 NAEP Reading by Item Type**

Item Type Grade 4 Reading	Number of Items	Item Type Grade 8 Reading	Number of Items
Total	170	Total	178
Multiple-Choice	108	Multiple-Choice	122
Short Constructed-Response	55	Short Constructed-Response	49
Extended Constructed-Response	7	Extended Constructed-Response	7

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, national Assessment of Educational Progress (NAEP), various years, 1990-2005 Mathematics Assessments.

**Table 2b. Percentage of NAEP Reading Items by Grade and Context for Reading**

Grade	Context for Reading		
	For Literary Experience	For Information	To Perform a Task
4	55	45	No Scale
Actual	51	49	
8	40	40	20
Actual	29	40	31

**Table 2c. Projected Distribution of Student Time by Grade and Aspect for Reading**

Grade	Aspect of Reading		
	Forming a General Understanding and Developing Interpretation (%)	Making Reader/Text Connections (%)	Examining Content and Structure (%)
4	60	15	25
Actual	68	14	17
8	55	15	30
Actual	59	17	24

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading.

**Chart 3. Item Distributions by Content Domains in Mathematics by State**

State	Assessment Program	Grade	Content Domain	Format	Test Item N. (max pt.)	% (pt.)
Louisiana	LA Educational Assessment Program (LEAP) <i>Mathematics</i>	4	Number and Number Relations	MC	30 no calculator	
			Algebra	MC	30 with calculator	(33)
			Measurement	SA	3 with calculator	(4)
			Geometry		max 72 pts	(8)
			Data Analysis, Probability, and Discrete Math			(14)
			Patterns, Relationships, and Functions			(18)
			8	Number and Number Relations	MC	30 no calculator
		Algebra	MC	30 with calculator	(21)	
		Measurement	SA	4 with calculator	(12)	
		Geometry		max 76 pts	(17)	
		Data Analysis, Probability, and Discrete Math			(16)	
		Patterns, Relationships, and Functions			(21)	
					(13)	

State	Assessment Program	Grade	Content Domain	Format	Test Item	
					N. (max pt.)	% (pt.)
Missouri	MO Assessment Program (MAP) <i>Mathematics</i>	4	Number Sense	MC	18; 1 (22)	37 (29)
			Geometric, Spatial Sense, and Measurement	CR	8; 3 (14)	21 (18)
			Data Analysis, Probability, and Statistics		4; 3 (12)	13 (16)
			Patterns and Relationships		1; 5 (11)	11 (14)
			Mathematical System and Number Theory		1; 4 (9)	10 (12)
			Discrete Mathematics		0; 4 (8)	8 (11)
			Total		52 (76)	
		8	Number Sense	MC	13; 2 (19)	29 (25)
			Geometric, Spatial Sense, and Measurement	CR	8; 3 (14)	21 (19)
			Data Analysis, Probability, and Statistics		4; 3 (10)	14 (13)
			Patterns and Relationships		4; 4 (12)	16 (16)
			Mathematical System and Number Theory		1; 4 (9)	10 (12)
			Discrete Mathematics		1; 4 (11)	10 (15)
			Total		51 (75)	

State	Assessment Program	Grade	Content Domain	Format	Test Item N. (max pt.)	% (pt.)
New Jersey	NJ Assessment of Knowledge and Skills (NJ ASK) <i>Mathematics</i>	4		MC	MC; OE (max pt)	
			Number Sense and Numerical Operation	OE	7; 2 (13)	27 (30)
			Geometry and Measurement		7; 1 (10)	24 (23)
			Patterns and Algebra		7; 1 (10)	24 (23)
			Data analysis, Probability, and Discrete Mathematics		7; 1 (10)	24 (23)
			Total			33 (43)
	Grade Eight Proficiency Assessment (NJ GEPA) <i>Mathematics</i>	8		MC	MC; OE (max pt)	
			Number Sense and Number Operation	OE	6; 2 (12)	22 (25)
			Geometry and Measurement		9; 1 (12)	28 (25)
			Patterns and Algebra		9; 1 (12)	28 (25)
Data analysis, Probability, and Discrete Mathematics				6; 2 (12)	22 (25)	
		Total			36 (48)	

State	Assessment Program	Grade	Content Domain	Format	Test Item N. (max pt.)	%	
West Virginia	WV Educational Standards Test (WESTEST) <i>Mathematics</i>	4	Numbers and Operations	MC; CR	20 (22)	38 (35)	
			Algebra		7 (9)	13 (15)	
			Geometry		10 (12)	19 (19)	
			Measurement		8 (10)	15 (15)	
			Data Analysis and Probability		7 (9)	13 (16)	
			Total			52 (62)	
		8	Numbers and Operations			9 (9)	17 (15)
			Algebra			15 (19)	29 (31)
			Geometry			10 (12)	19 (19)
			Measurement			7 (9)	13 (15)
Data Analysis and Probability				11 (13)	21 (21)		
	Total			52 (62)			

State	Assessment Program	Grade	Content Domain	Test Item		
				Format	N. (max pt.)	% (pt.)
South Carolina	Palmetto Achievement Challenge Tests (PACT)	4	Number and Operations	MC; CR	MC; CR (max pt.)	
			Algebra		9 (9); 1 (3)	26 (27)
			Geometry		7 (7)	18 (16)
			Measurement		6 (6); 1 (3)	18 (20)
			Data Analysis and Probability		7 (7); 1 (2)	21 (20)
		Total	5 (5); 1 (3)	16 (18)		
		8	MC; CR	MC; CR (max pt.)		
		Number and Operations		12 (12); 2 (5)	24 (24)	
		Algebra		14 (14); 2 (4)	28 (25)	
		Geometry		13 (13); 1 (3)	24 (22)	
Measurement	6 (6); 1 (2)	12 (11)				
Data Analysis and Probability	10 (10); 1 (3)	19 (18)				
Total		58 (72)				

**Table 3. Target Percentage of Distributions of NAEP Items by Content Area<sup>1</sup>**

<b>2005 Mathematics</b> <i>Content Category</i>	<b>Grade 4</b>			<b>Grade 8</b>		
	<i>N.</i> <sup>2</sup>	<i>Target %</i>	<i>Actual %</i>	<i>N.</i>	<i>Target %</i>	<i>Actual %</i>
<i>Number Sense, Properties, and Operations</i>		40			20	
<i>Measurement</i>		20			15	
<i>Geometry and Spatial Sense</i>		15			20	
<i>Data Analysis, Statistics, and Probability</i>		10			15	
<i>Algebra and Functions</i>		15			30	
<b>Total</b>	<b>181</b>	<b>100</b>		<b>197</b>	<b>100</b>	

<sup>1</sup> Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), various years, 1990-2005 Mathematics Assessments.

<sup>2</sup> The number of items in each content category by test booklet is not available. The total number is for multiple test booklets. Sources: U.S. Department of Education, Institute of Education Statistics, National Assessment Progress (NAEP), 1990, 1992, 1996, 2000, and 2003 Mathematics Assessments.

**Table 4. Inclusion Rate for the 2005 NAEP**

State	Grade	Reading		Mathematics	
		Identified	Assessed	Identified	Assessed
California	4	39	34	39	35
	8	28	25	28	25
Delaware	4	20	7	20	12
	8	17	6	18	7
Louisiana	4	24	10	24	20
	8	16	8	15	11
Missouri	4	17	9	18	16
	8	16	8	15	11
New Jersey	4	18	12	18	15
	8	18	13	18	14
West Virginia	4	18	13	20	17
	8	18	11	17	14

Sources of NAEP data: Percentage of all students identified as students with disabilities and/or English language learners, excluded, and assessed, when accommodations were permitted, grade 4 [grade 8] public schools: By state, various years, 1998-2005.

**Table 5. Ranking of Cut Scores Stringency and Performance On NAEP for Six States**

Category Item Description	Grade	Sample States Used in the Study							
		CA	DE	LA	MO	NJ	SC	WV	
Ranking for Reading <sup>1</sup>	4	7	n/a	16	n/a	20	2	25	
	8	5	24	11	n/a	14	2	30	
Ranking for Mathematics	4	n/a	n/a	16	5	20	4	27	
	8	n/a	12	25	1	15	2	33	
Percent of Proficiency on the 2007 NAEP	4	22.9	33.8	20.4	31.8	43.1	25.8	27.8	
Reading (%) <sup>2</sup>	8	21.5	30.3	19.4	31	39	24.6	22.9	
Percent of Proficiency on the 2007 NAEP	4	29.7	40	24.4	38.4	51.8	35.9	32.6	
Mathematics (%)	8	23.9	31.3	19	29.9	40.4	31.9	18.5	
NAEP Scale Score Change in Reading	4	2.9	1.1	2.7	-1.5	5.6	-1	-4	
2003 to 2007	8	0.3	0	0.2	-3.9	2.4	0.7	-4.6	
NAEP Scale Score Change in Mathematics	4	2.6	5.9	3.8	4.6	9.8	1.3	5.6	
2003 to 2007	8	3.5	5.8	6.1	1.9	7.2	4.2	0.7	
Poverty Gap: National School Lunch Program									
Non-Eligible vs. Eligible on 2007 NAEP Reading	4	30.3	18.5	25.3	21.5	27.3	26.6	18.9	
Poverty Gap Change 2003 to 2007	4	0.6	-1.2	-4	-2.6	-3.2	0.7	3.4	
Non-Eligible vs. Eligible on 2007 NAEP Mathematics	8	26.2	20.2	19.5	24.5	31.3	25	19.2	
Poverty Gap Change 2003 to 2007	8	-3.6	-3.7	-4.6	1.6	-3.2	0.9	0.7	

Data sources 1: Mapping 2005 State Proficiency Standards onto the NAEP Scales (Braun and Qian, 2007)

Data sources 2: Quality Counts 2008 - A Special Supplement to Education Week's, January 9, 2008

