STATE-BY-STATE COMPARISONS OF STUDENT ACHIEVEMENT: THE DEFINITION OF THE CONTENT DOMAIN FOR ASSESSMENT

Robert L. Linn

CSE Report No. 275

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

1987

The project presented, or reported herein, was performed pursuant to a Grant from the Office of Educational Research and Improvement/Department of Education (OERI/ED). However, the opinions expressed herein do not necessarily reflect the position or policy of the OERI/ED and no official endorsement by the OERI/ED should be inferred.

State-By-State Comparisons of Student Achievement: The Definition of the Content Domain for Assessment

Robert L. Linn

University of Colorado, Boulder

Twenty years ago when the National Assessment of Educational Progress (NAEP) was being designed, care was taken to ensure that the data would not allow comparisons among individual states or localities. There were a variety of reasons for this decision, including considerations of cost, political viability, and concerns about the likely misuse of state average scores on the assessment. Today, however, the lack of information at the level of individual states has been judged to be the most serious weakness of NAEP by the blue ribbon panel that was constituted to review NAEP and make recommendations about its future (Alexander-James, 1987).

The NAEP Study Group, which was chaired by Governor Lamar Alexander and directed by H. Thomas James, identified the development of state-by-state comparative data as its number one priority. The Study Group reasoned that most "important decisions in education are made at the state or local level, and accountability for performance is vested at those levels" (p. 4). They also implied that the decision makers at the state or local level would benefit from comparative information, but did not explicitly state how such information would be used to make better educational decisions.

The Study Group considered some of the concerns that, in the past, had led to a decision to prevent the use of NAEP for purposes of making state-by-state comparisons, but concluded that the "concerns are less important now than they were previously, and that most can be readily accommodated within a redesigned national assessment" (Alexander-James, 1987, p. 5). Having thus dismissed the

objections to state-to-state comparisons under the heading "previous concerns about comparisons", the Study Group was ready to give its most important recommendation.

> The single most important change recommended by the Study Group is that the assessment collect representative data on achievement in each of the fifty states and the District of Columbia. Today state and local school administrators are encountering a rising public demand for thorough information on the quality of their schools, allowing comparison with data from other states and districts and with their own historical records. Responding to calls for greater accountability and for substantive school improvements, state officials have increasingly turned to the national assessment for assistance (pp. 11-12).

The movement toward state-by-state comparisons, of course, did not begin with the Alexander-James Study Group. Rather, the Study Group endorsed a position that had already garnered considerable support from policy makers and the public during the past five years. Comparisons of schools, school districts, and states in terms of student achievement have become popular and are seen by many as a means of fostering competition and stimulating improvement. As Gallup and Clark (1987) have indicated, the demand for comparisons and the view that they will foster improvements in education comes from the highest levels of government and has considerable popular support.

> In his 1984 State of the Union Address, President Reagan asserted, 'Just as more incentives are needed within our schools, greater competition is needed among our schools. Without standards and competition there can be no champions, no records broken, no excellence - in education or any other walk of life.' The public agrees. Seventy percent favor reporting the results of achievement tests by state and by school, so that comparisons can be made. The public feels that such comparisons would serve as incentives to local public schools, whether the results showed higher or lower scores for local students (Gallup & Clark, 1987, p. 19).

The call for state-by-state comparisons by the Alexander-James Study Group was also consistent with the desires of a number of educational policy makers. The movement toward state-by-state comparisons was encouraged earlier by the U.S. Department of Education and by the Council of Chief State School Officers (CCSSO).

CCSSO has provided considerable support for the idea of state-by-state comparisons during the past three years since it adopted a position paper encouraging states to develop comparable measures of student achievement in reading, mathematics, English, science, and social studies. The subsequent establishment of the State Education Assessment Center by CCSSO with the support of the Center for Statistics and the Mott Foundation and the activities of the Assessment Center and CCSSO since that time have given greater strength to the movement toward making state-by-state comparisons a reality. With support from the U.S. Department of Education and the National Science Foundation, CCSSO is now in process of forming a consortium of educators that will develop specific recommendations for the first state-by-state assessment of student achievement in mathematics.

As Ramsay Selden (1986a), the director of the State Education Assessment Center, has noted, any approach that is taken to the development of a system that will yield state-by-state comparisons of student achievement will raise "profound issues in educational measurement" (p. 2). Selden went on to discuss some of those issues and highlighted the need to deal with issues of validity. The focus of this paper is on a limited set of issues related to the validity of the assessment system. More specifically, the purpose of this paper is to review issues concerning the definition of the domain of content to be covered in the assessment and the relationship of the definition and score reporting systems to the validity of inferences that are based on state-by-state comparisons.

VALIDITY

As with any use of tests, the most fundamental measurement issue in the development of an assessment system that will provide state-by-state comparisons is the validity of the inferences that will be made from the scores. To date, however, relatively little serious attention has been given to the questions of

validity of a NAEP based state-by-state comparison system, or for that matter, any other system other than the seriously flawed use of college admissions test scores as indicators of the educational quality in a state.

Although not couched in terms of validity, the primary concern that was raised in the National Academy of Education's review committee commentary on the Alexander-James report is fundamentally an issue of validity. The Review Committee (National Academy of Education, 1987, p. 59) summarized its reservations about the recommendation that NAEP be redesigned to provide state-by-state comparisons as follows.

> We are concerned about the emphasis in the Alexander-James report on state-by-state comparisons of average test scores. Many factors influence the relative rankings of states, districts, and schools. Simple comparisons are ripe for abuse and are unlikely to inform meaningful school improvement efforts.

As is clearly implied by the above statement, the Review Committee's concern applies not only to the proposed state-by-state comparisons using NAEP but to the use of average test scores for other units such as individual school buildings or school districts. The concern is not limited to the use of NAEP. It would apply equally well to the use of other assessment devices or tests. The concern is clearly with the inferences that the Review Committee anticipated would be made from the test data. The validity of those inferences will depend on a wide variety of factors, such as the degree of standardization of the rules for inclusion and exclusion of students in the assessment, the specific sampling procedures, and the administration procedures. One of the important factors that will influence the validity of the inferences drawn from the comparisons, however, is the adequacy of the content coverage of the assessment. Issues regarding the validity of the content assessment will be considered in the remainder of this paper.

CONTENT DOMAIN

It is one thing to agree that the assessment should cover the "core content areas (reading, writing, and literacy; mathematics, science, and technology; history, geography, and civics)" (Alexander-James, 1987, p. 12), but quite another to agree that a particular set of topics in, say, history, much less that a specific set of items, should be included on the assessment that is to be used to compare states. It is also much easier to achieve agreement that "the assessment instruments should examine acquisition of pertinent 'higher-order' skills as well as basic skills, knowledge, and concepts" (Alexander-James, 1987, p. 8), than it is to gain consensus that a given exercise is a fair assessment of higher-order thinking skills. Many of the issues that arise when a school or district selects a test are also relevant at the state level. Among these are the issues of the breadth of the coverage, the match between what is taught and what is tested, the number and specificity of the scores that are reported, and the familiarity of the assessment procedures that are used.

Breadth of Coverage and the Match with What is Taught.

Since the issues of breadth of coverage and that of the degree to which the assessment matches the curriculum and what is actually taught in classroom are closely related, they will be considered together. One approach to the determination of the content to be included in an assessment would be to require a consensus among all states that a given topic or assessment exercise is appropriate to the state's instructional goals for students at a given point in their educational program. As Selden (1986a) has noted, the consensus about a "common body of knowledge could be conceived as a 'least common set' -- that content which is pursued to some degree by schools in each [state], but excluding anything which all states cannot be presumed to be teaching or emphasizing. Alternatively, it could be conceived as an 'optimal set', around which consensus can be reached, but which may not reflect everything some states

are pursuing, and which may include some items that some states may not be pursuing or emphasizing" (p. 7). To these two alternatives could be added, at least in theory, an "inclusive set", that content that is judged to be appropriate by one or more states.

Although the "inclusive set" is apt to be too unwieldy in practice, it illustrates an end of a continuum that is anchored at the other end by the "least common set". On the surface, the least common set appears the fairest approach. It would not hold a state accountable for students learning content that was not expected to be taught in its schools by a given grade level. However, as will be discussed in some detail below, the least common set approach can be faulted on several accounts, including that of fairness.

The issue of where along the continuum between the least common and the inclusive sets an assessment should be placed is not unique to the present context. It long has been an important issue in the use of tests in program and curriculum evaluation (e.g., Burstein, 1981; Cronbach, 1963; Walker & Schaffarick, 1974; Wargo & Green, 1978). If a test does not measure the outcomes that correspond to important program goals, the evaluation will surely be considered unfair. The judgment that the evaluation is unfair takes on additional force when multiple programs are compared and the tests used to measure the educational outcomes of the programs appear to match the goals of one program better than another.

The latter point is clearly illustrated by the controversy that surrounded the Follow Through evaluation. Follow Through was a massive federal experiment that pitted twenty-two early education models against each other over the course of ten years. The model programs varied considerably in their stated goals but were evaluated using a common set of outcome measures. Between-model differences were found on some of the subtests of the Metropolitan Achievement

Test (MAT) (Stebbins, St.Pierre, Proper, Anderson, & Cerva, 1977). The differences occurred on subtests that the evaluators classified as "basic skills" and favored models that were classified as emphasizing basic skills over models that were classified as having a "cognitive-conceptual" emphasis or an "affective-cognitive" emphasis. Press accounts of the evaluation presented the message that education that emphasizes the basics yields the best results.

Because of the potential importance of the Follow Through evaluation, the Ford Foundation sponsored a comprehensive third-party review of the evaluation. The review resulted in a devastating critique that faulted the evaluation on numerous grounds (House, Glass, McLean, & Walker, 1978). Of most relevance to the present discussion, however, is the House, et al. critique of the measurement of the program outcomes and the characterization of those outcomes. Their analysis led them to conclude that "the outcome measures assess very few of the models' goals and strongly favor models that concentrate on teaching mechanical skills" (House, et al, 1978, p. 156).

Although not strictly a question of test content, the format of the test items and administrative procedures can also have implications for the results of an assessment. Even apparently trivial changes in item format, such as the presentation of addition problems horizontally rather than vertically, have been found to effect the scores that children obtain (Alderman, Swinton, & Braswell, 1979). More importantly, the outcome of an assessment can be affected by the match between the format used to ask question on the test and the format used when students practice the skill in the instructional program and the amount of practice that they have with similar tests (Alderman, et al., 1979; Cooley & Leinhardt, 1980; Roberts, 1980).

The match between what is taught and what is tested can have a substantial effect on the performance on tests. The closer the match and the more the test questions tap rote memory, the larger the likely effects. Indeed, two of the

most compelling examples involve the choice of words for tests of spelling or for the vocabulary used to assess beginning reading. Hopkins and Wilkerson (1965) compared four forms of the California Spelling Test to the course of study guide used in California. Because the forms varied in the degree to which they matched the study guide, knowledge of only those words that were in the curriculum study guide would yield scores that differed by as much as 2.1 grade equivalent units depending on which of the four forms was used. As would be expected, the California students were much more likely to correctly respond to words that were in the curriculum than words that were not.

Bianchini's (1978) analysis of the remarkable increase in the percentile rank of the California state median reading achievement test score for first grade students between 1970 and 1971 provides another example of the dramatic effect that the degree of match between what is taught and what is tested can have on tests scores. Over the course of that single year, the median score for first grade students throughout the state rose from the 38th to the 50th percentile. As Bianchini's analyses suggests, however, the huge increase had more to do with the fact that the test that was used to measure reading achievement was different in 1971 than it was in 1970, than to any dramatic increase in the quality of education provided to first grade children. Bianchini found that 55% of the vocabulary on the test that was used in 1971 was included in the state's first grade readers, whereas only 19% of the vocabulary on the test used the previous year was included in the readers.

Results such as those reported by Bianchini (1978), Hopkins and Wilkerson (1965), and others (e.g., Cooley & Leinhardt, 1980; Leinhardt, 1983; Leinhardt & Seewald, 1981) might lead one to believe that "least common" set approach is necessary to avoid unfair comparisons. However, the solution is not that simple. To begin with, the fact that two programs both teach children to add

fractions, for example, does not imply that both programs give that skill the same priority or spend an equal amount of time teaching it. If the children at one school were drilled extensively on the addition of fractions with little attention given to other arithmetic operations or to mathematics concepts, while children at a second school spent some, but much less, time on that skill while spending considerably more on other skills and on concepts and problem solving, a test that only measured the addition of fractions would hardly be considered fair. As in the case of the Follow Through evaluation, the test would strongly favor the first school because it lacked more comprehensive coverage of the skills and concepts that were emphasized at the second school. While such extremes are unlikely to be encountered in practice, even at the level of individual schools much less at the level of entire states, the example illustrates the fact that the use of the least common set will tend to favor those who emphasize the skills and concepts contained in that set at the expense of those that are not included in the set.

No matter what process is used to define the domain of content, it must include knowledge, skills, and concepts that educators, policy makers, and the general public consider important. This is part of the reason that the Alexander-James (1987) report emphasized that the assessment should include measures of higher-order thinking skills which the report defined to include "recognizing a problem's general structure, defining goals, isolating the information relevant to problem solutions, ... evaluating the merits of arguments, ... reasoning, analyzing, explaining, and finding analogies" (p. 15). Such a list does not appear to be compatible with the least common set approach to defining the domain to be assessed for purposes of state-by-state comparisons.

Experience suggests that when a consensus is required for a topic or problem type to be included on an assessment, challenging, but important, topics

and problems are often excluded (Pandey, 1987). The common set becomes the "least common denominator". Minimums are more readily agreed to as standards against which one will be held accountable than are maximums. Furthermore, as Brown (1987, p. 50) has argued, "It is the policy makers' job to establish minimum standards" and although he acknowledges that minimum standards have utility in education as well as other areas, he also warns that "The problem with a policy that is minimum-oriented, however, is that it requires a free market or some other such device to push it toward excellence" (1987, p. 50).

The Alexander-James (1987) list of higher-order thinking skills would push the assessment beyond a minimum set of basic skills that would be likely to define the least common set to a broader set of goals. Inasmuch as there is general agreement that higher-order thinking skills of the type envisioned by the Alexander-James study group should be taught, the list is in keeping with what Selden (1986b) has referred to as the "optimal consensus" approach wherein the content of the assessment would be defined to include content for which a consensus can be reached that given content knowledge and skills should be taught. The idea of this approach is that it would allow the assessment to go beyond minimal objectives that are already pursued by all and thereby have a potentially broadening influence on the curriculum rather than a narrowing influence that is apt to be associated with least-common-set approach.

Increased breadth of coverage requires large numbers of test questions, more than could reasonably be administered to each student. This suggests that it will be important to maintain some form of matrix sampling procedure whereby each student in the sample responds to only a fraction of the total set of test questions that are administered in a given assessment. This approach is commonly used on state assessments and has been a part of NAEP since its beginning. Hence, the need for a means of administering more items than can be

taken by any given student poses no new problems.

Increased depth of assessment of the type envisioned by the Alexander-James Study Group, on the other hand, poses a greater challenge. It is much easier to write questions that assess simple recall of factual knowledge or the routine application of rules that can be memorized and applied with little, if any, understanding of the underlying principles (e.g., the division of one fraction by another), than it is to write questions that assess a deeper understanding of principles; the ability to integrate, evaluate, and apply information; and the ability to solve ill-formed problems where problem identification and representation, the generation of hypotheses, planning and the exploration of solution strategies are of central importance.

A number of authors (e.g., Frederiksen, 1984; Newmann, 1987; Romberg, 1985), have suggested that the assessment of higher-order skills such as those listed above will require that assessments go beyond traditional multiple-choice items. Romberg (1985, p. 5), for example, argued that the multiple-choice item format "imposes inherent limitations on how much one can tell about how students respond to mathematical questions" such as a student's ability "to produce a sustained deductive proof." In a similar vein, Newmann (1987, p. 1) has argued that the production of discourse (i.e.,, "a narrative, argument, explanation, or analysis") and not merely the recognition of the best choice on a multiple-choice question is critical to adequate assessment is social studies. And, Frederiksen (1984, p. 199) has argued that multiple-choice items typically present only well-structured problems, that is, ones that "are clearly stated, all the information needed to solve the problem is available in the problem or (presumably in the head of the student) and an algorithm exists that guarantees a correct solution if properly applied." He went on to argue convincingly, however, that most important problems, both in and out of school, are ill-structured. They do not have a single correct answer, much less provide all the

information needed to solve the problem.

Although research has generally shown that item format makes little, if any, difference when "existing multiple-choice tests [are compared] with their free-answer counterparts" (Frederiksen, 1984, p. 199), the converse is not true. That is, When we begin with existing free-response tests designed to measure more complex cognitive problem solving skills, different results are found (Frederiksen, 1984, p. 199). The formulating hypotheses test developed by Frederiksen and Ward (1978; Ward, Frederiksen, & Carlson, 198Ø) demonstrates that a paper-and-pencil test with open-ended responses can be used effectively to assess certain higher-order skills that are crucial to scientific reasoning and which are not readily tapped in a multiple-choice format.

If the challenge to assess higher-order thinking skills that was presented by the Alexander-James Study Group is taken seriously, it is incumbent upon the designers of the assessments to explore a wide range of approaches to assessment. The use of open-ended response problems and attempts such as the use of hands-on problems in the recent NAEP assessment in science need to be expanded. The assessment needs to go beyond the least-common-set approach and beyond minimum basic skills. The assessment needs to encourage both greater breadth and greater depth of content coverage.

Number and Specificity of Scores

If these goals are to be realized, the assessment will need to have a content domain with broadly defined limits and emphasize more than simple factual knowledge. As Anderson (1986) has noted, such an assessment is apt to measure several dimensions of achievement within each subject area and raise questions about the nature and number of scores to be reported.

Cronbach (e.g., 1963, 1971) has long argued that for purposes of evaluation, a comprehensive array of measures should be sought. "An ideal

evaluation might include measures of all the types of proficiency that might reasonably be desired in the area in question, not just the selected outcomes to which ... [a particular] curriculum directs substantial attention" (Cronbach, 1963, p. 680). The assessment needs to provide a basis for identifying areas that are judged to be important but that students are not learning, whether or not the poor learning is the result of lack of exposure. Furthermore, for purposes of making decisions about the curriculum or program of instruction, the test results need to be reported separately for each of the specific areas of proficiency, and not merely combined into a single overall score.

The latter point runs counter to the goal of having a simple score card that will allow the ranking of states along a single dimension. However, Cronbach's rationale for maintaining separate scores is compelling.

> If the original test or battery is a composite covering various types of content or various objectives, it implicitly weights those elements, either by the number of items allocated to each or by the way the score is calculated. Such a weighting cannot satisfy decision makers who hold values unlike those of the test developer. Consequently, an ideally suitable battery for evaluation purposes will include separate measures of all outcomes the users of the information consider important ... Reporting separate scores allows for the application of various systems of values. It also enables the investigator to examine the nature of any weaknesses in the program. (emphasis in the original) (Cronbach, 1971, p. 460).

The use of a single composite score not only forces an implicit set of values on the outcome of the assessment and prevents those who hold different values from seeing the results from that alternate framework, but the composite may sometimes be insensitive to differences between the educational systems that are being compared (Airasian & Madaus, 1983; Madaus, Airasian, & Kellaghan, 1980). In other instances, and of even greater concern, the composite may favor a system with an emphasis that happens to match the content that the composite weights most heavily.

The latter problem is illustrated by the results of Walker and Schaffarzick's (1974) review of twenty-six studies that compared students who

had been exposed to a given subject matter using either "traditional" or "innovative" curriculum materials and then tested with one or more measures of achievement. Their review provides strong evidence that "different curricula are associated with different patterns of achievement" (emphasis in the original) (p. 97). Whether the results of the studies reviewed favored the "traditional" or the "innovative" curriculum was largely determined by the content of the tests. "Students using each curriculum do better than their fellow students on tests which include items not covered at all in the other curriculum or given less emphasis there" (Walker & Schaffarzick, 1974, p. 97). If a single global score were used to compare the alternative curricula an outcome of no difference, one favoring the traditional curriculum, or one favoring the innovative curriculum could be readily achieved according to the relative weighting given to the test content favoring each.

The need to report multiple scores corresponding to clearly and specifically defined content areas is convincingly demonstrated by recent experience with tests that are customized to the specifications of a state or local district. The need for multiple scores can also be demonstrated from recent experience with the NAEP assessments in literature and U. S. history. In both instances it is evident that a single total score can conceal specific areas of strength and weakness. Furthermore, the relative standing of a given state, region, or other aggregate of students can be greatly influenced by the number of items in a total score that happen to be associated with specific content areas.

The dangers of relying exclusively on a total score when the number of items with specific content areas varies have recently become apparent in the use of item banks designed to provide users with norm-referenced interpretations of results. In the past, if a state or district wanted to compare the achievement of its students to a national norm, it had to administer a norm-referenced test.

If the state or district also wanted to obtain test results on a test designed to match locally defined objectives, a second test administration was generally required since the standardized test would not match the locally defined objectives as closely as desired. Recently, however, test publishers have begun offering an option of creating a "customized" test that consists of items selected according to locally specified objectives, but from which norm-referenced scores are also produced.

Customized tests are the result of increased use of item response theory by publishers in their test development and scaling process. One of the features of item response theory that makes it especially appealing is the promise that, once the theory has been used to calibrate a pool of test items, any set of items from that pool can be used to place the performance of test takers on a common scale (see, for example, Hambleton & Swaminathan, 1984, chapter 12). Thus, according to the theory, any set of previously calibrated items could be selected by a state or district to be included among those on its customized test and the resulting test scores could still be placed on the same scale as the published version of the standardized test for which national norms are available.

The quality of the norm-referenced scores that a state or district obtains for its customized test depends on several factors, including (1) the adequacy of the item response theory model for the set of items in the calibrated item pool, (2) the number of calibrated items selected for the customized test, (3) the statistical characteristics of the items selected from the item pool, and (4) the degree to which items selected for the customized test match the content coverage of the published version of the test for which the norms are available. Recent experience with a major customized test, the Kentucky Essential Skills Test (KEST), suggests that the last of these four considerations can be of

critical importance (Linn, 1986; Yen, Green, & Burkett, 1987).

Kentucky administered the KEST to essentially all eligible students in the state in grades K through 12 for the first time in 1985. The 1985 KEST was a customized test, containing, among other items, items that were selected from the CTB/McGraw-Hill item pool. That pool includes items from the Comprehensive Tests of Basic Skills (CTBS), Forms U and V, items from the California Achievement Tests, Forms C and D, and previously unpublished items. Since all items are calibrated to the CTBS scale, a test that had previously been administered statewide in Kentucky, it was possible to obtain estimates of performance on the CTBS scale from the administration of the KEST. When the KEST results were obtained in 1985, however, at least two major anomalies were observed. The most notable and troublesome of these was a precipitous increase in the grade 5 mathematics test performance.

In 1982, 1983, and 1984, when the CTBS was administered statewide to fifth grade students, the state mean normal curve equivalent (NCE) scores in mathematics ranged from 50.4 to 54.8. In 1985, however, the mean NCE for grade 5 mathematics based on the KEST was 66.3. Thus, on the NCE scale, which has a standard deviation of 21 for the national norm group, the state mean increased in a single year by slightly over a half of the national norm group standard deviation. A review of the KEST and the calibration of the items in the item pool from which it was constructed did not suggest that the application of item response theory was any more problematic than in many other widely accepted applications. However, it was evident that the grade 5 mathematics results on the 1985 KEST could not be meaningfully compared to the earlier CTBS results (Linn, 1986).

The lack of comparability between the KEST and CTBS grade 5 mathematics tests is most plausibly explained by differences in the proportion of items on the KEST and the CTBS that are classified into specific content categories.

The proportions of KEST and CTBS grade 5 mathematics items by content category were as follows (Linn, 1986).

Content Category	CTBS Proportion	KEST Proportion
Numeration	.42	.27
Number Theory	.Ø3	.13
Measurement	.16	.11
Geometry	.1Ø	.2Ø
Number Sentences	.19	.2Ø
Problem Solving	.1Ø	.Ø9

As was demonstrated by Yen, Green, and Burkett (1987), systematic differences as a function of content category between local and national estimates of item response theory difficulty parameters are sometimes found. Such differences can lead to misleading global score results when content coverage changes. "Content equivalence between customized and normed tests is essential if the customized test is to be NRT-equivalent and norm-valid" (Yen, Green, & Burkett, 1987, p. 13). Separate reporting by specific content categories, however, is needed in order to identify areas of strong and weak performance and to make value judgments about the importance of changes in scores on the global score.

The final example illustrating the importance of multiple scores corresponding to specific content categories comes from the recent NAEP results in literature and U. S. history (Applebee, Langer, & Mullis, 1987). Both the literature and the U. S. history item sets met the usual criteria for deciding if a unidimensional item response theory model is appropriate. Hence, single global performance scores were estimated for each of the two broad content domains.

Despite the apparent simplicity for each content area, however, substantial differences that could be meaningfully interpreted were found for content specific subsets of items as a function of region of the country, gender, and

race/ethnicity. For example, even though the performance of black test takers was well below that of whites on the bulk of the literature and U. S. history items, blacks outperformed whites on questions asking about black leaders or black literature. Black test takers also did better than whites on several of the questions dealing with slavery and civil rights. Similarly, though women outperformed men on the overall literature scale, men did better on "items focusing on strong male literary characters" (Applebee, et al., 1987, p. 3), such as Robin Hood, King Arthur, Samson, and Captain Ahab. Although the Southeast region of the country scored well below the northeast on the overall literature scale, the converse was true on the 15 items dealing with Biblical characters and stories.

The above examples illustrate two points that are of great potential importance in any future state-by-state comparisons of student achievement. First, the rank order on a single global score is apt to depend on the particular weighting of the content categories. Based on the KEST results, one might reasonably expect, for example, that Kentucky would have appeared better on a grade 5 test with heavy emphasis on numeration than on one that emphasized another content category such as number theory or geometry. Second, a single global score can also conceal educationally important information about strengths and weaknesses in the curriculum.

The need to focus on multiple content specific outcomes has been recognized within the context of state assessments by Bock and his colleagues (Bock & Mislevy, 1987; Bock, Mislevy, & Woodson, 1982; Mislevy, 1983). For purposes of informing curriculum planners, assessment information needs to be provided for highly specific content areas which Bock, Mislevy, and Woodson (1982) called "indivisible curricular elements". These are "item domains that are sufficiently homogeneous with respect to content that all the items in a given

domain would be similarly affected by changes in curricular emphasis" (Mislevy, 1983, p. 273).

SUMMARY AND CONCLUSIONS

It has been argued that the choice of content for a state-by-state comparisons will be one of many factors that will have a substantial influence on the validity of inferences that may be drawn from a state-by-state assessment system. Based on considerable experience in the use of tests in the evaluation of alternative educational programs, it was concluded that there are great disadvantages to an approach that focuses only on content and skills that are thought to be taught in a given grade in all states. Such a "least-common set" approach would be likely to give a relative advantage to states that narrow their focus to only that least common set. The approach is more likely to narrow than to broaden the curriculum.

Ideally, the domain for assessment would include separate measures of the full range of outcomes that are considered important by any of the states. The multiple measures would enable states to identify strengths and weaknesses and not just obtain a ranking on a global score card. The more inclusive set would encourage a broadening rather than a narrowing of the curriculum by calling attention to wide range of outcomes.

Despite the desirability of having multiple scores corresponding to "indivisible curricular elements" for purposes of identifying strengths and weaknesses and planning changes in the curriculum, such scores clearly will not satisfy the demand for a overall number in reading or a single score for mathematics. Global scores will certainly need to be produced, in part, because the amount of information would be too overwhelming for many of its intended uses if it were only reported at the level of indivisible curricular element level, and, in part, because there is a desire, as Ambach (1987) has noted, for a score card. Global scores can, and undoubtedly, will be produced. The

argument here is not that such scores should not be produced, but that the ability to disaggregate the results to more specific content areas should be maintained. The disaggregated scores are needed to interpret the overall results and plan improvement.

References

Airasian, P. W. & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. Journal of Educational Measurement, 20, 103-118.

Alderman, E. L., Swinton, S. S., & Braswell, J. S (1979). Assessing basic arithmetic skills and understanding across curricula: Computer-assisted instructional and compensatory education. Journal of Children's Mathematical Behavior, 2, 3-28.

Alexander-James Study Group. (1987). The nation's report card: Improving the assessment of student achievement. Cambridge, MA: National Academy of Education.

Ambach, G. M. (1987). Testing and educational quality. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1985) Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Anderson, P. S. (1986). Beyond the wall chart: Issues for states. Technical Report. Portland, OR: Northwest Regional Education Laboratory.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1987). Literature and U. S. history: The instructional experience and factual knowledge of high school juniors. Princeton, NJ: Educational Testing Service.

Bianchini, J. D. (1978). Achievement tests and differential norms. In M. J. Wargo & D. R. Green (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill.

Bock, R. D. & Mislevy, R. J. (1986). Comprehensive educational assessment for the states: The duplex design. Technical Report, Chicago. IL: National Opinion Research Corporation.

Bock, R. D., Mislevy, R. J., & Woodson, C. E. M. (1982). The next stage of educational assessment, Educational Researcher, 11, 4-11, 16.

Brown, R. (1987). Who is accountable for 'thoughtfulness'? Phi Delta Kappan, 69, No. 1, 49-52.

Burstein, L. (1981). Investigating social programs when individuals belong to a variety of groups over time. CSE Technical Report # 173. Los Angeles, CA: UCLA Center for the Study of Evaluation.

Cooley, W. W. & Leinhardt, G. (1980). The instructional dimensions study. Educational Evaluation and Policy Analysis, 2, 7-25.

Cronbach, L. J. (1963). Evaluation of course improvement. Teachers College Record, 64, 672-683.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement, Second edition. Washington, DC: American Council on Education. pp 443-507.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

Frederiksen, N. & Ward, W. C. (1978). Measures for the study of creativity in scientific problem solving. Applied Psychological Measurement, 2, 1-24.

Gallup, A. M. & Clark, D. L. (1987). The 19th annual Gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan, 69, No. 1, 17-30.

Hambleton, R. K. & Swaminathan, H. (1984). Item response theory: Principles and applications. Boston, MA: Khuwer-Nijhoff.

Hopkins, K. D. & Wilkerson, C. J. (1965). Differential content validity: The California Spelling Test. Educational and Psychological Measurement, 25, 413-419.

House, E. R., Glass, G. V., McLean, L. D., & Walker, D. F. (1978). No simple

answer: Critique of the Follow Through evaluation. Harvard Educational Review, 48, 128-160.

Leinhardt, G. (1983). Overlap: Testing whether it's taught. In G. F. Madaus (Ed.), The courts, validity, and minimum competency testing. Hingham, MA: Kluwer Nijhoff.

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

Linn, R. L. (1986). Norm-referenced score estimates from the Kentucky Essential Skills Test. In Center for the Study of Testing, Evaluation, and Educational Policy (George Madaus, Principal Investigator). An evaluation of the Kentucky Essential Skills Test. Chestnut Hill, MA: Boston College.

Madaus, G. F., Airasian, P. W., & Kellaghan, T. (1980). School effectiveness: A reassessment of the evidence. New York: McGraw-Hill.

Mislevy, R. (1983). Item response models for grouped data. Journal of Educational Measurement, 8, 271-288.

National Academy of Education. (1987). Review Committee commentary of The Nation's Report Card. Cambridge, MA: National Academy of Education.

Newmann, F. M. (1987). The assessment of discourse in social studies. Unpublished manuscript, Madison, WI: University of Wisconsin, National Center for Effective Secondary Schools, September.

Pandey, T. (1987). Issues of curriculum alignment. Paper presented at a conference on Approaches to Subject Matter Assessment hosted by the Center for Evaluation, Standards, and Student Testing, UCLA, December, 1987.

Roberts, A. O. H. (1980). Practice effect and test-wiseness. Mountain View, CA: RMC Research Corporation.

Romberg, T. A. (1985). The content validity, for School mathematics in the U.S., of the mathematics subscores and items for the Second International Mathematics Study. Paper prepared for the Committee on National Statistics,

National Academy of Sciences.

Selden, R. (1986a). Some classical measurement issues confronting the development of state-by-state assessment of student achievement. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA, April.

Selden, R. (1986b). White paper. Strategies and issues in the development of comparable indicators for measuring student achievement. State Education Assessment Center, Council of Chief State School Officers., April 30.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B., & Cerva, T. R. (1977). Educational as experimentation: A planned variation model, Volume IV-A, An evaluation of Follow Through. Cambridge, MA: Abt Associates, Inc.

Ward, W. C., Frederiksen, N., & Carlson, S. (1980). Construct validtiy of free-response and multiple-choice versions of a test. Journal of Educational Measurement, 17, 11-29.

Wargo, M. J. & Green, D. R., (Eds.). (1978). Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill.

Yen, W., Green, D. R., & Burkett, G. R. (1987). Valid normative information from customized achievement tests. Educational Measurement: Issues and Practice, 6, 7-13.