

2002

# Teachers' Use of High-Stakes Test Results to Improve Instruction: A Review of Literature

James H. McMillan
*Virginia Commonwealth University*, jhmcmill@vcu.ede

Susan P. McKelvey
*Virginia Commonwealth University*, smckelvey@sunllc.com

# Teachers' Use of High-Stakes Test Results To Improve Instruction:

# A Review of the Literature

**James H. McMillan, Professor**
**Educational Studies**
**Virginia Commonwealth University**

**Susam McKelvy, MERC Research Fellow**
**Virginia Commonwealth University**

**May 2002**

**Teachers' Use of High-Stakes Test Results To Improve Instruction: A Review of the Literature**

Across the United States many school districts and state departments of education have embraced high-stakes testing for their public schools and students. This has led to widespread school reform since many students have not reached the standards that the states and/or districts have set. Teachers increasingly focus on making sure that their students have enough knowledge to pass these tests, and some states and districts have made passing scores a requirement for graduation. Many states and districts have invested considerable time aligning their standardized tests with the objectives of the curriculum.

One of the consequences of high-stakes testing is that teachers have become more accountable for what they do in their classrooms. Of particular relevance to this review is the professional development that teachers engage in to understand and use students' standardized test scores. The emphasis is on helping teachers review the scores and data of these tests so as to make informed decisions regarding instructional practices. This review will examine literature that has addressed the use of standardized test scores by teachers to improve instruction and student learning. The guidelines, principles, and suggestions are presented by level of general application, beginning with broad, general principles of test use and ending with practices specific to MERC school divisions. Figure 1 provides an overview of the major categories that will be covered.

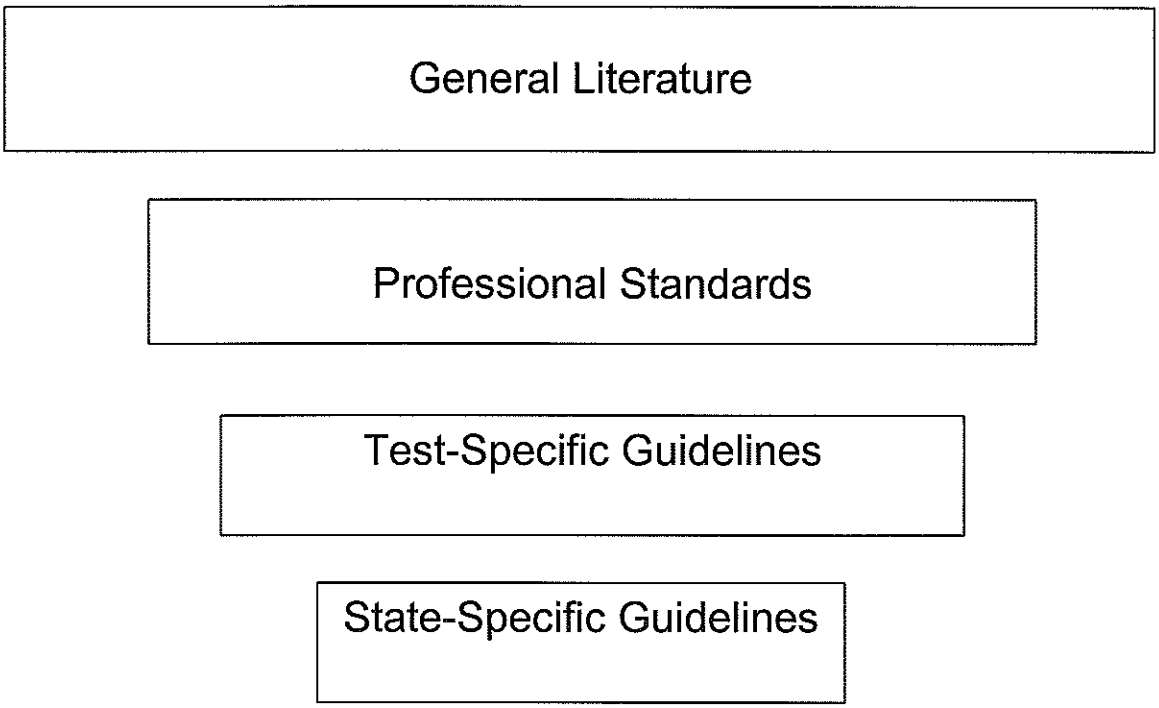



Figure 1. Categories Reviewed by Level of Generality

## General Principles

When examining the use of standardized tests – high-stakes tests, in particular – there are several principles that test users should consider. First and foremost, standardized tests should only be one factor when making placement, instructional or promotional decisions about students (Carter, 2000; McMillan, 2001; Mehrens & Lehman, 1987; Payne, 1997). Other tools, such as classroom assessments, are needed to verify suggested results. Payne (1997) believes that past performance and other experience should be considered in conjunction with the test scores. Furthermore, he states that, "Information about skills and knowledge already acquired and developed is immensely helpful in designing future educational programs for individual students, groups, or classes" (p. 438).

Before gathering data on students and schools, one needs to understand intentions for data use. In this new high-stakes assessment movement, the data are used for many reasons, including: identifying strengths and weaknesses of the curriculum or of students (Mehrens & Lehman, 1987; Payne, 1997); improving instruction (McMillan, 2001); and predicting how well students will do in the future (McMillan, 2001; Payne, 1997). Other possible uses consist of ability grouping of students and determining if remedial instruction for particular students is necessary (McMillan, 2001; Mehrens & Lehman, 1987). Mehrens and Lehman (1987) further suggest that these data can reveal the effectiveness of the methods of instruction.

Even though test data have multiple uses, Mehrens and Lehman (1987) believe that every standardized test has its strengths and weaknesses. Also, to make the best decisions concerning data use, test users must acknowledge key factors that may affect test scores.

For instance, student attendance or levels of motivation may influence the scores. In addition, a child's home life, health or academic ability may have an effect. If teachers do not consider such factors when interpreting test data, they may reach inaccurate conclusions about the students (Payne, 1997). Creighton (2001) discusses "data-driven decision making" (p. 9), which calls for locating patterns in the data over several years. For example, if males do better in math than females consistently, over time, teachers and administrators may decide to make changes in the curriculum based on this information.

Researchers have suggested other general rules or principles to consider when analyzing or interpreting test data. For example, use of item analysis is appropriate to identify strengths and weaknesses, but not to assign grades (Mehrens, 1987). Because reporting results by item can lead to an over abundance of data, users need to select what will be most useful. School districts ultimately determine how the data will be disaggregated, and they should select only a few variables for review over several years (Wahlstrom, 1997). In addition, the standards should be well aligned with district curriculum, instruction, and assessment (Holcomb, 1999). Last, as mentioned previously, student performance should not be judged on the basis of a single test (Linn, 2000).

Teachers should have a basic understanding of testing and statistics when interpreting test scores for instructional decision-making. This includes some technical concepts, such as reliability and validity. Reliability and validity play important roles in the interpretation of test scores. The American Educational Research Association, the American Psychological Association and the National Council on Measurement, in their *Standards for Educational and Psychological Testing* (1999), define validity as "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (p. 9). They

further stress that, "The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretation" (p.9). The clear intent is that there are valid or invalid inferences and uses associated with test results, not valid or invalid tests. Reliability refers to "the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups" (p. 25). Like validity, reliability is concerned with the interpretation of the scores.

Reliability and validity are applicable to criterion-referenced, norm-referenced, and standards-referenced tests. Criterion-referenced tests rely on pre-established levels of performance for interpretation. They can take several forms, such as specifying "the probability that an examinee's level of tested knowledge or skill is adequate to perform successfully in some other setting" (*Standards*, 1999, p. 50). Norm-referenced testing allows the comparison of an individual or a group of individuals to a defined population using percentiles and class averages, and also enables identification of strengths and weaknesses. Criterion- and norm-referenced tests are different because of the nature of the interpretation. Criterion-referenced is "performance with regard to the domain" (Payne, p. 17), a measurement that gives percentages correct or placement in relation to established standards, while the norm-referenced measurement shows comparisons (Payne, 1997). Most high-stakes tests have predetermined cut scores to indicate different levels of proficiency. This information may be somewhat useful for teachers, but it is necessary to fully understand what performance at each level means. Often, such cut-score results give only a very general indication of student performance or ability, and more specific information is needed to design instruction based on the scores.

The *Standards* (1999) provides a glossary with definitions of testing terms that are useful to establish a common understanding among all who use test scores. Some of the more relevant terms pertaining to our review include the following:

**absolute score interpretation** - The meaning of a test score for an individual or an average score for a defined group, indicating an individual's or group's level of performance in some defined criterion domain.

**criterion-referenced test** - A test that allows its users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others. Examples of criterion-referenced interpretations include comparison to cut scores, interpretations based on expectancy tables, and domain-referenced score interpretations.

**cut score** - A specified point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point.

**derived score** - A score to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

**error of measurement** - The difference between an observed score and the corresponding true score or proficiency.

**fairness** - In testing, the principle that every test taker should be assessed in an equitable way.

**high-stakes test** -A test used to provide results that have important, direct consequences for examinees, programs, or institutions involved in the testing.

**local norms** - Norms by which test scores are referred to a specific, limited *reference population* of particular interest to the test user (e.g., locale, organization, or institution); local norms are not intended as representative of populations beyond that setting.

**norm-referenced test interpretation** - A score interpretation based on a comparison of a test taker's performance to the performance of other people in a specified *reference population*.

**percentile** - The score on a test below which a given percentage of scores fall.

**percentile rank** - Most commonly, the percentage of scores in a specified distribution that fall below the point at which a given score lies. Sometimes

the percentage is defined to include scores that fall at the point; sometimes the percentage is defined to include half of the scores at the point.

**reliability** - The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group.

**scaling** - The process of creating a scale or a scale score. Scaling may enhance test score interpretation by placing scores from different tests or test forms onto a common scale or by producing scale scores designed to support criterion-referenced or norm-referenced score interpretations.

**test user** - The person(s) or agency responsible for the choice and administration of a test, for the interpretation of test scores produced in a given context, and for any decisions or actions that are based, in part, on test scores.

**validity** - The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of tests. (pp. 171-184)

Recently, five associations that serve the needs of classroom teachers, school principals, and district superintendents commissioned a group of nationally recognized experts in assessment, curriculum, and instruction to develop nine "requirements" of responsible high-stakes assessment to improve both learning and accountability. Particular emphasis was placed on how accountability tests could be designed and implemented to help teachers do a better job in the classroom, to be instructionally supportive. Five national associations endorsed the requirements (AASA, NAESP, NASSP, NEA, and NMSA). Several of the requirements have direct implications for teacher use of test scores, as summarized in the following (*Building Tests to Support Instruction and Accountability: A Guide for Policymakers*, 2001):

| Requirement | Implications for Teachers |
|---|---|
| 1. "A state's content standards must be prioritized to support effective instruction and assessment." (p. 4) | How do administrators help teachers to understand what content is emphasized on the test? |
| 2. "A state's high priority content standards must be clearly and thoroughly described so that the knowledge and skills students needed to demonstrate competence are evident." (p.5) | To what extent do teachers understand the nature of the knowledge and skills tests? |
| 3. "The results of a state's assessment of high-priority content standards should be reported standard-by-standard for each student, school and district." (p. 5) | How should teachers pull items together for a standard or other reporting category? How do teachers utilize scores comparing different students, schools, or districts? |
| 4. "A state must ensure that educators receive professional development focused on how to optimize children's learning based on the results of instructionally supportive assessments." (p. 5) | What is the nature and delivery of professional development that will ensure teachers' effective use of test scores? |
| 5. "A state should secure evidence that supports the ongoing improvement of its state assessments to ensure those assessments are (a) appropriate for the accountability purposes for which they are used, (b) appropriate for determining whether students have attained state standards, (c) appropriate for enhancing instruction, and (d) not the cause of negative consequences." (p. 5) | To what extent do the assessments enhance instruction? What are possible negative consequences of the assessments? |

The general literature on the use of standardized and standards-based test scores suggests the following guidelines:

- Verify standardized test scores with other data and information
- Recognize potential sources of error (e.g., student motivation)
- Identify patterns of student achievement over several years
- Avoid overwhelming teachers with data

- Disaggregate data
- Ensure teacher understanding of technical and statistical concepts (e.g., validity, reliability, and standard error of measurement)
- Prioritize standards to be stressed
- Provide professional development in appropriate test score interpretation and use

## Professional Standards

There are several organizations that have developed professional standards for the appropriate use of test score data. We will summarize those that have the greatest relevance for teachers' use of standardized and standards-based tests.

The Joint Committee on Testing Practices developed the *Code of Fair Testing Practices in Education* in 1988. The *Code* contains standards for test developers and users, with sections on interpreting scores and fairness that are relevant for teacher use of standards-based test results, as delineated in the following:

Test users should:

- interpret scores correctly;
- obtain information about the scale used for reporting scores, the characteristics of any norms or comparison group(s), and the limitations of the scores;
- interpret scores taking into account any major differences between the norms or comparison groups and the actual test takers;
- take into account any differences in test administration practices or familiarity with the specific questions in the test;
- avoid using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use;
- explain how any passing scores were set and gather evidence to support the appropriateness of the scores;

- obtain evidence to help show that the test is meeting its intended purpose;
- review the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available.

The *Standards for Teacher Competence in Educational Assessment of Students* (1990) were developed by the American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association. The third and fourth standards are related to interpretation and use of external tests. The standards are followed by a set of skills needed to provide for professionally sound uses of the test, as indicated in the following:

| Standard | Skills |
|---|---|
| Teachers should be skilled in administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods. | • Administer standardized achievement tests and interpret reported scores.<br>• Understand summary indexes, including measures of central tendency, dispersion, relationships, and errors of measurement.<br>• Analyze assessment results to determine student strengths and weaknesses. |
| Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and making recommendations for school improvement. | • Use accumulated assessment information to organize a sound instructional plan.<br>• Interpret results correctly according to established rules of validity.<br>• Use results from local, regional, state, and national assessments for educational improvement. |

The *Code of Professional Responsibilities in Educational Measurement* (1995) lists responsibilities for individuals who: (1) develop assessments; (2) market and sell assessments; (3) select assessments; (4) administer assessments; (5) score assessments; (6) interpret, use and communicate assessment results; (7) educate about assessments; (8) evaluate programs and conduct research on assessments. Responsibility six relates directly to

this review. It focuses on the need to make valid inferences and includes the following relevant standards:

> 6.2 Provide to those who receive assessment results information about the assessment, its purposes, its limitations, and its uses necessary for the proper interpretation of the results.
>
> 6.3 Provide to those who receive score reports an understandable written description of all reported scores, including proper interpretations and likely misinterpretations.
>
> 6.4 Communicate to appropriate audiences the results of the assessment in an understandable and timely manner, including proper interpretations and likely misinterpretations.
>
> 6.5 Evaluate and communicate the adequacy and appropriateness of any norms or standards used in the interpretation of assessment results.
>
> 6.6 Inform parties involved in the assessment process how assessment results may affect them.
>
> 6.7 Use multiple sources and types of relevant information about persons or programs whenever possible in making educational decisions.
>
> 6.8 Avoid making, and actively discourage others from making, inaccurate reports, unsubstantiated claims, inappropriate interpretations, or otherwise false and misleading statements about assessment results.
>
> 6.10 Report any apparent misuses of assessment information to those responsible for the assessment process.

The most comprehensive set of professional testing standards is set forth in the *Standards for Educational and Psychological Testing* (1999). These *Standards* were developed and approved by the American Psychological Association, the National Council on Measurement in Education, and the American Educational Research Association. The following standards, categorized by major chapters, relate to this review:

**Validity**

- A rationale should be presented for each recommended interpretation and use of test scores.
- Potential users should be cautioned about making unsupported interpretations.
- When interpretation of performance on specific items, or small subsets of items, is suggested, the rationale and relevant evidence in support of such interpretations should be provided. When interpretation of individual item responses is likely but is not recommended, the user should be warned against making such interpretations.

**Reliability**

- For each total score, sub-score, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement should be reported.

**Scales and Score Comparability**

- Test documents should provide test users with clear explanations of the meaning and intended interpretations of the derived scale scores, as well as their limitations.
- If there is sound reason to believe that specific misinterpretations of a score scale are likely, test users should be explicitly forewarned.
- When raw scores are intended to be directly interpretable, their meanings, intended interpretations, and limitations should be described and justified.

**Responsibilities of Test Users**

- Prior to the adoption and use of a published test, the test user should study and evaluate the materials provided by the test developer. Of particular importance are those that summarize the test's purposes and discuss the score interpretations for which validity and reliability data are available.

- The test user should have a clear rationale for the intended uses of a test or evaluation procedure in terms of its validity and contribution to the assessment and decision-making process.

- Test users should be alert to potential misinterpretations of test scores and to possible unintended consequences of test use; users should take steps to minimize or avoid foreseeable misinterpretations and unintended negative consequences.

- In educational, clinical, and counseling settings, a test taker's score should not be interpreted in isolation; collateral information that may lead to alternative explanations for the examinee's test performance should be considered.

**Educational Testing and Assessment**

- When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both tested and targeted domains should be described in sufficient detail so their relationship can be evaluated. The analyses should make explicit those aspects of the target domain that the test represents as well as those aspects that it fails to represent.

- In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision.

- In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.

- In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of these differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

The National Board for Professional Teaching Standards has published *What Teachers Should Know and Be Able To Do* (2001). For each subject and subject level, the Board includes a standard specifically for assessment practices. Every section states that teachers should use a variety of appropriate assessments, and that the "accomplished" teacher uses the information from multiple sources to make determinations about student learning. Most of the assessment standards also state that the teachers should improve instructional practices and allow the students to reflect on their strengths and weaknesses.

Early this fall, 2002, the Joint Committee on Standards for Educational Evaluation, housed at Western Michigan University, will issue *Student Evaluation Standards*. These standards have been in development for several years with participation from major educational organizations (e.g., AASA, AERA, NCME, ASCD, NEA, and AFT). The

standards provide agreed upon principles for conducting and interpreting tests and other means of assessment to evaluate student knowledge and skills. Of the 29 standards, five have direct application to the use of standardized test scores:

| Standard | Description |
|---|---|
| Evaluator Qualifications | Teachers and others who evaluate students should have the necessary knowledge and skills so that the evaluations are carried out competently and the results can be used with confidence. |
| Evaluation Support | Adequate time and resources should be provided for student evaluations, so that evaluations can be effectively planned, implemented, interpreted, and communicated. |
| Validity Orientation | Student evaluation should be developed and implemented, so that the interpretations made about the performance of a student are valid and not open to misinterpretation. |
| Context Analysis | Student and contextual variables that influence performance should be identified and considered, so that each student's performance can be validly interpreted. |
| Defensible Information | The adequacy of information gathered should be ensured, so that good decisions are possible and can be defended/justified. |
| Analysis of Quantitative Information | Quantitative information from student evaluations should be systematically and accurately analyzed, so that the purposes and the uses of the evaluation are effectively served. |

The relevant standards that have been summarized are synthesized in the following list to show professional competencies and practices that provide a foundation for appropriate use of standardized and standards-based test scores.

- Provide teachers with information about the purpose, limitations, and uses of test scores
- Provide descriptions of the meanings of all scores
- Identify appropriate and inappropriate interpretations and negative consequences
- Use multiple sources of information in making instructional decisions
- Provide explanations of scales used
- Avoid using tests for purposes not specifically recommended

- Test scores should only be used by those with adequate training
- Each intended use of test scores should have a clear and defensible rationale
- Clarify the extent to which the test samples targeted domains
- The degree of measurement error should be reported with all test score results
- Group differences in test results should be accompanied by relevant contextual information
- Summary scores for groups of students should be accompanied by measures of variability

## Test-Specific Suggestions

*Iowa Test of Basic Skills(ITBS)*

Riverside Publishing Company publishes the ITBS *Interpretive Guide for Teachers and Counselors* to help teachers interpret students' scores. The company stresses that ITBS test results should supplement teachers' observations and classroom assessments. It is suggested that teachers initially take some steps to establish the integrity of the score information, including a check to make sure all students are included and examining whether results generally match with classroom data. Any "outlier" score should be identified and considered when computing averages. The publisher suggests administering the tests during the fall so that teachers can use results for immediate instructional modifications. While examining the test data, teachers should focus on subscale scores as well as total scores. For example, vocabulary and comprehension subscales fall under Reading, and those scores are totaled to determine a general Reading score.

When using the Individual Performance Profile (IPP), strengths and weaknesses can be readily determined for particular students. The profile disaggregates the student's scores even further to allow for better analysis of his or her performance. On the other hand, the Student Criterion-Referenced Skills Analysis actually shows the teacher which skill would be a strength or a weakness by a designated "+" sign or a "-" sign in front of the particular skill.

Riverside Publishing also recommends looking at the averages for particular classes and subject areas on the Profile Chart for Averages to analyze groups of students, as opposed to individual students.

Teachers can examine test scores to tailor instructional practices for particular students. The company provides a List Report of Student Scores, Individual Performance Profile, and a Student Criterion-Referenced Skills Analysis. The ITBS developers also list suggestions for teachers once they have determined a weakness.

## State-Specific Suggestions

*Maryland*

Because of the national trend with high-stakes testing, several states have had years of experience in helping teachers use the results. Specifically, Maryland, Kentucky, Connecticut and Texas will be examined.

The Maryland assessments, created by the Maryland State Department of Education (MSDE), are called the Maryland School Performance Assessment Program (MSAP). They are for third, fifth and eighth grades, with the Maryland Functional Testing Program for ninth and eleventh grades. The MSDE has been providing the public with state, local and specific school results since 1991. Some data are used for descriptions of characteristics, with no standards, while the other data are used with standards of "satisfactory" or "excellent" categories. Only school-level scores are currently reported.

When downloading the data from the website, the user has the choice of viewing the data by state, locality or school. One can choose to view a line graph that compares how many students received a score of "satisfactory" or "excellent" on a specific test since 1993.

The website also states that at least 70% of students must receive a score of "excellent", and at least 25% of students must receive a score of "satisfactory" before the school meets its standard. In addition, the site provides race and ethnicity data, gender data, "Students Receiving Special Services" data, and the number of students tested and absent.

When analyzing the data, MSDE suggests using graphs (including bar and pie), with comparison to the state or school system, and school. On the website, information can be found tailored to school administrators, which assists them with data preparation and presentation to their school staff. The website also includes tips for educators on how to prepare students for each area of the test. It has a section on benchmarking schools, so that the teachers and administrators can see how their school compares to other similar schools.

*Kentucky*

On the Kentucky Department of Education website, a portion exists that contains a guide for educators and parents to help them interpret the data from their Kentucky Core Content Test (KCCT). The state tests the students in the core content areas, differentiating between four levels of scores: Novice, Apprentice, Proficient and Distinguished. Each school needs to reach a score of Proficient, and the test has cut points that delineate each level.

The Kentucky State Department of Education provides each school with its own Kentucky Performance Report (KPR). Included in this report is trend data for each content area, dating back to 1999; disaggregated data based on many factors, such as race/ethnicity and gender; mean scores in reading, math, science and social studies; and many other factors to illustrate performance for each subject area in the elementary, middle and high schools. In addition, each school receives the Kentucky Core Content Report (KCCR), which provides

more detailed data on student performance on sub-domain levels. The interpretive guide also provides a glossary of specific terms necessary for analyzing data.

Kentucky has also made available training modules for individuals visiting the site, even though the training audience is administrators and not parents or teachers. Module Three, for example, relates to using the data to make decisions. During an activity included in this module, the trainer provides sample data and the participants analyze the data to find areas of concern. They then develop "Next Steps" and answer the following questions: What are the next steps for using and sharing the Performance Level Descriptions in our school/district? How do we begin to embed the Performance Level descriptions in our work to improve curriculum and instruction? The participants use these questions to guide them in completing a table to outline how and when teacher training will occur.

The site provides several different ways to view the KCCT test results: Adobe Acrobat Reader, Microsoft Word, a generic word processor, or a database file. This feature is provided for elementary, middle and high school results. For the elementary level, the test is administered to grades four and five; for middle school, grades seven and eight; and in the high school, grades ten and twelve. Once downloaded, the results are listed with average scores for the district level and each school level for the last three years.

*Connecticut*

The state of Connecticut administers two tests, one called the Connecticut Mastery Test (CMT) for grade four, and the other called Connecticut Academic Performance Test (CAPT) for grade ten. The CMT tests math, reading and writing, while the CAPT tests four content areas: Math, Reading across the Disciplines, Writing across the Disciplines and Science.

For the CAPT, students earn overall scale scores for each content area, ranging from 100-400. To meet state goal standards, students need to have a score ranging from 250-261, depending on the test. The state has four levels, with the top level termed Goal Level and the bottom Intervention Level. The website provides an overview of the results on a statewide level, disaggregated results (by race, gender, etc.), an overall summary for each school, district and the entire state, and a year-by-year comparison. The site also gives the percentage of students above the goal, separated by school.

For the CMT, the Connecticut Department of Education has two publications: *Language Arts Handbook 2001* and *Mathematics Handbook.* Both provide instructional tips and practice tests for teachers to administer. These handbooks are available for the public to download from the website.

Connecticut has an *Interpretive Guide* on the Web for the CAPT, but not for the CMT. The first paragraph states that teachers must be aware of any other variables influencing a student's performance on a standardized test. The Guide also stipulates that the scale scores assist in looking at a student's performance over several years; however, scale scores cannot be compared across disciplines. With scores ranging from 100-400, score reports are given to parents with an explanation of the scores and what the student demonstrates with a particular score. Neither the *Guide* nor the website has any information for teachers to help them alter their instruction based on the score interpretations.

*Texas*

Like the other three states, Texas also has helpful information on its website. Texas has four assessments: Texas Assessment of Academic Skills (TAAS), State Developed Alternative Assessments (SDAA), end-of-course examinations, and Reading Proficiency

Tests in English (RPTE). TAAS is administered for reading and math to grades three through eight and at the high school level. It also measures writing for grades four through eight, as well as in high school. In addition, the science and social studies versions are for eighth graders, and a Spanish version is administered to grades three through six (to Spanish-speaking students). Students must pass all of the TAAS tests given at the high school level to receive their diplomas. The TAAS will be replaced with the Texas Assessment of Knowledge and Skills in 2003. The end-of-course tests are given statewide as well, but only in high school for specific courses, such as Algebra I, Biology, English II, and U.S. History. Again, the students must pass these tests to graduate. The RPTE is for students learning English as a second language. These students take this test from grades three through twelve, until they receive an advanced score. The SDAA is specifically for special education students in grades three through eight.

Texas has an interpretive guide for all four tests on the site as well. For the TAAS and the end-of-course tests, students can earn a scale score from 400-2400, with a minimum passing score of 1500. The students can also earn a "Mastered All Objectives" when they show that they have mastered every objective for a subject area test. In addition, they can earn "Academic Recognition" as long as they answer at least 95% of the multiple-choice items correctly and receive a certain score (4) on the writing test.

Texas Education Agency provides a fairly extensive resource to help teachers and administrators interpret and use the results. There is a heavy emphasis on using item and objective results as indicators in which further diagnosis is warranted. In discussing the use of the scores for placement decisions, they stress that the results should be used in conjunction with other performance data. In recognition of the domain-referenced nature of the tests, they

indicate that generalizations may only be made to the specific content domain represented by the objectives that are tested. Year to year comparisons by objective should be made with caution. There is also a mechanism provided for having a "level playing field" for comparing different districts.

*Virginia*

Virginia implemented the Standards of Learning (SOLs) in 1995, and began testing students on the SOL in 1998. The SOL developers considered the standards that students should meet at specific grade levels and created tests for the four core subject areas. The possible range of scores is 0 to 600 for each SOL test. To be proficient in each content area, a student must score at least 400; to meet the advanced level, a student must score at least 500. Both are scaled scores, and one content area cannot be compared to another (R. Triscari, personal communication, January 29, 2002). At least 70% of students must pass the SOLs in all four core subject areas for the school to have full accreditation (Regulations, 2000). Harcourt-Brace prepares the data for each school division in Virginia.

James Heywood, the Director of Elementary School Instructional Services for the Virginia Department of Education offered presentations called *Analyzing Your SOL Test Data: Improving Instruction Through Data-Based Decisions*. Part I, ("Are You Aligned?") stresses collaboration among staff, as well as consistency in all schools. Heywood also discusses the SOL Blueprint, which outlines the percentage of questions for each standard on each test. For example, he shows a slide that displays part of the Blueprint for the math categories. For grade five, 16% of the questions are dedicated to "Number/Number Sense". He suggests that in the classroom teachers align the time they spend on a category with the percentage of questions on the SOL. Part II of his presentation is titled "SOL Data Analysis".

Heywood states that teachers should look at PALS data (for elementary school teachers), grade five data (for middle school teachers), and grade eight data (for high school teachers) so that they will have a context in which to place SOL results. He says that teachers should not just examine the mean scores, as they can be misleading given the individual score frequency distribution. The teachers need to target students individually by using each student's data. He also recommends analyzing data in reporting categories by comparing students who pass the test (but score below 30) with students who fail the test (and score below 30). He offers tips to administrators so that they can assist their teachers. First, he recommends giving teachers the resources they need (SOL Blueprint, curriculum, etc.). Second, he suggests offering staff development for specific categories. Third, Heywood stresses analyzing data teacher by teacher. Lastly, he recommends having grade and/or department meetings to discuss findings. The final portion of his presentation focuses on reading problems in the elementary schools (Heywood, 2001).

One of the critical factors in effective use of the SOL scores is alignment, the degree of correspondence among the Standards of Learning, Standards of Learning tests, instruction, and classroom assessments. Determining appropriate alignment requires professional judgment to determine whether instruction and classroom assessments match with the SOL and SOL tests. At a superficial level, these judgments are informal and focus on the content of what is taught and tested. At a more sophisticated level, judgments can be made about the nature of the knowledge that is emphasized and the nature of the cognitive process that is required of students. These two dimensions, knowledge and cognitive process, are examined at length in a recently published revision of Bloom's taxonomy (Anderson and Krathwohl, 2001). Table 1 summarizes these two dimensions.

Table 1
Types of Knowledge and Cognitive Processes

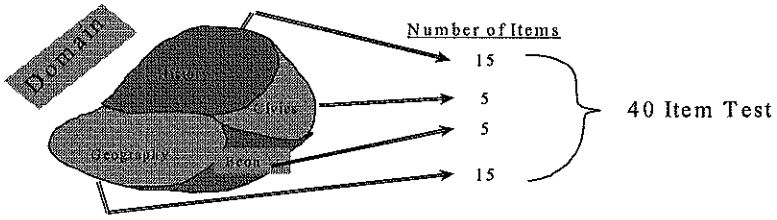
| | Description | Examples |
|---|---|---|
| **Types of Knowledge** | | |
| Factual | Basic elements students must know to be acquainted with a discipline or solve problems in it. | Technical vocabulary, major natural resources, musical symbols, letters, numbers. |
| Conceptual (principles) | Interrelationships among the basic elements within a larger structure that enable them to function together. | Forms of business ownership, law of supply and demand, theory of evolution, structure of congress. |
| Procedural | How to do something, methods of inquiry, and criteria for using skills, algorithms, techniques, and methods. | Painting in watercolors, whole-number division algorithm, using interviewing techniques, scientific method. |
| Meta-cognitive | Knowledge of cognition in general as well as awareness of knowledge of one's own cognition. | Knowledge of outlining as a means of capturing the structure of a unit or subject matter; knowledge of the cognitive demands of different tasks; awareness of cognitive strengths and weaknesses. |
| **Type of Cognitive Process** | | |
| Remember | Retrieve relevant knowledge from long-term memory. | Recognizing or recalling dates of important events in US history. |
| Understand | Construct meaning from instructional messages. Clarifying, interpreting, classifying, summarizing, comparing, and explaining. | Changing from one form of representation to another; drawing logical conclusions; determining a specific example of a concept or principle. |
| Apply | Carry out or use a procedure in a given situation. | Applying a procedure to a familiar task, such as determining the circumference of a circle; using laws in new situations. |
| Analyze | Break material into its constituent parts and determine how the parts relate to one another and to an overall structure or | Distinguishing relevant from irrelevant parts; determining how elements fit or function within a structure; determining a point of view |

| | | |
|---|---|---|
| | purpose. | or bias. |
| Evaluate | Make judgments based on criteria and standards. | Detecting inconsistencies or fallacies within a process, procedure or product; judging a product or procedure in relation to established criteria. |
| Create | Put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure. | Coming up with alternative hypotheses; devising a procedure to accomplish as task; inventing a product or procedure. |

This new taxonomy is helpful because it incorporates more recent cognitive theories and research on learning and student motivation. The implications of the taxonomy for alignment are readily apparent when such alignments are made on the basis of type of knowledge and cognitive process. That is, an examination of the type of knowledge represented by the SOL reporting categories and illustrated by specific test items, as well as the nature of the cognitive processes required, is needed to determine whether instruction and classroom assessments, which can be evaluated in the same way, match one another. This is a more sophisticated type of alignment than that in which there is only matches based on content.

## Domain-Referenced Nature of SOL Tests

The SOL tests are domain-referenced. The term "domain-referenced" is appropriate because items are randomly selected from item banks to be representative of a larger domain of knowledge as determined by the reporting categories and standards. In Figure 2 the concept of domain-referencing is illustrated with the 5$^{th}$ grade social studies test.




5th Grade Social Studies

Figure 2. Illustration of Domain Referenced Nature of 5th Grade SOL Social Studies Test

The smallest unit of knowledge is what is represented by the reporting categories. It is not an individual standard that may be tested, often with a single item. To provide a fair evaluation of a specific standard several items testing that standard would need to be included.

## Summary

It is clear from a review of the literature on teachers' use of standardized test scores that there has been little empirical research to document, or even suggest best practices. On the other hand, it is also clear that there are many established professional standards that can guide policy and practice. There is certainly extensive use of SOL test scores by teachers. What remains to be investigated is the extent to which suggested practices are being implemented and the extent to which there is a direct relationship between receipt of the scores and instruction.

The general literature, professional standards, test publishers, and current state use of scores suggest a number of *potential* "best practices" that can be used as the basis for further study. The following is a summary of these potential best practices and principles:

- Assure teacher knowledge of important technical aspects of testing, such as validity, reliability, standard error of measurement, and scale scores
- Assure teacher knowledge of SOL test blueprints, reporting categories, and domain-referenced nature of SOL tests
- Verify test score results with other information; use results as an indicator that warrants further diagnosis
- Recognize potential sources of error when interpreting results; report standard errors of measurement
- Identify patterns of information over several years
- Use graphic displays to better understand test score results
- Participate in professional development in test score interpretation
- Identify appropriate and inappropriate score interpretations

- Justify each intended use of scores with a clear and defensible rationale
- Focus interpretation on group average reporting category scores; generalize to the domain represented by the standards that are tested
- Make comparisons with caution
- Do not compare percentage correct for different SOL test items

## References

*Building tests to support instruction and accountability: A guide for policymakers.* (2001). Commission on Instructionally Supportive Assessment.

*Code of fair testing practices in education* (1988). Washington, DC: American Psychological Association [Electronic Version]. Retrieved December 18, 2001 from http://wwwcsteep.bc.edu/CTESTWEB/documents/related/codeoffairtest.html.

*Code of professional responsibilities in educational measurement* (1995) National Council on Measurement in Education [Electronic Version]. Retrieved November 7, 2001, from http://www.natd.org/Code_of_Professional_Responsibilities.html.

Connecticut Department of Education (n.d.). *Students and testing.* Retrieved November 7, 2001, from http://www.csde.state.ct.us/public/der/s-t/index.htm

Heywood, J. (2001). Analyzing your SOL test data: improving instruction through data-based decisions. *Presentation given to Richmond City (Richmond, VA) school administrators, November 2001.*

Hoover, H., et. al. (1993). *Iowa tests of basic skills : Interpretive guide for teachers and counselors.* Chicago, IL: The Riverside Publishing Company.

Kentucky Department of Education (n.d.). *School report card.* Retrieved November 7, 2001, from http://www.kde.state.ky.us/oaa/implement/school_report_card/info/

Linn, R. (2000). *Assessments and accountability.* Based on a paper presented in Educational Researcher (v.29 n2 p 4-16).

Maryland State Department of Education (n.d.). *Maryland school performance report.* Retrieved November 7, 2001, from http://msp.msde.state.md.us/

McMillan, J. (2001). *Classroom assessment: Principles and practice for effective instruction (2nd ed.).* Needham Heights, MA: Allyn & Bacon. 390-391.

Mehrens, W., & Lehmann, I. (1987). *Using standardized tests in education (4th ed.).* New York: Longman. 307-312.

National Center for Education Statistics (2001.). *The nation's report card* [Electronic Version]. Retrieved February 6, 2002, from http://nces.ed.gov/nationsreportcard/naepdata/.

Payne, D. (1997). *Applied educational assessment.* Belmont, CA: Wadsworth Publishing Company.

Regulations Establishing Standards for Accrediting Public Schools in Virginia, 8 VAC 20-131-10 et. seq. (Adopted by VA Board of Education July 28, 2000) (codified on September 28, 2000).

*What teachers should know and be able to do* (2001). National Board for Professional Teaching Standards [Electronic Version]. Retrieved November 7, 2001, from http://www.nbpts.org/standards/know_do/policy.html.

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

*Standards for Teacher Competence in Educational Assessment of Students* (1990). American Federation of Teachers, National Council on Measurement in Education, National Education Association.

Texas Education Agency (n.d.). *Welcome to the student assessment division.* Retrieved November 7, 2001, from http://www.tea.state.tx.us/student.assessment