What Teachers Need to Know about Assessment

Edited by

Lawrence M. Rudner
William D. Schafer

Contributions by

Linda A. Bond
Carol Boston
Amy Brualdi
Lucy Calkins
Robert B. Frary
Jerard Kehoe
Jon A. Leydens
Diane Loulou
James McMillian
William A. Mehrens
Kate Montgomery
Barbara M Moskal
Bonnie Potts
Donna Santman
Christopher Tienken
Michael Wilson

Published by the National Education Association

Note:

The National Education Association commissioned this book with the intention of making available to educators a single wide-spectrum work on student assessment, containing both original new selections and edited reprints of important articles collected in the Educational Resources Information Center (ERIC) Clearinghouse on Asse3ssment and Evaluation (ERIC/AE).

ERIC staff members' time for this project was sponsored in part with funding from the Office of Educational Research and Improvement (OERI), U.S. Department of Education, under contract ED99CO0032.

The opinions expressed in this publication should not be construed as representintg the policy or positions of the National Education Association, nor do they necessarily reflect the positions or policy of OERI or the U.S. Department of Education. The materials are intended to be discussion documents for educators who are concerned with specialized interests of the profession.

Library of Congress Cataloguing-in-Publication data.

Rudner, Lawrence
What teachers need to know about assessment/Lawrence M. Rudner,
William D. Schafer. p. cm. - (Student assessment series)
ISBN 0-8106-2074-X
1. Educational tests and measurements - United States. 2.
Examinations - United States - Design and construction. 3.
Examinations - United States - Scoring. I. Schafer, William D. II. Title.
III. Series

LB 3051 .R815 2002
371.26'2 - dc21 2002016607

Preface

One doesn't need to look very far to see how important testing and assessment have become in American education. On almost a daily basis, we see test results or testing ideas in the popular press or in memos from the state or district Superintendent's office.

Testing is more than accountability. It can be a means to improve education, itself. Standardized tests and large-scale assessments can be used, and are being used, to encourage teaching of the skills prescribed by state and local agencies. A critical component of instruction, various forms of teacher assessment permeate everyday classroom activity. Paper and pencil tests provide formal feedback with regard to what has and has not been learned. The routine asking of questions and the scoring of projects and activities in the classroom are other forms of assessment that strike at the heart of instruction. Teachers' need for information is commensurate with the pace of their instructional decision making, which is probably more intense than in any other profession.

Teachers today, perhaps more so than ever before, have a need to be knowledgeable consumers of test information, constructors of assessment instruments and protocols, and even teachers about testing. Few courses and textbooks exist to help meet this need and there are very few materials designed specifically for teachers in the classroom.

The goal of this book is to help you become a knowledgeable user of teacher-constructed and district/state sponsored assessments. You will learn

- fundamental concepts common to all assessments;
- essential classroom assessment concepts.
- useful concepts and issues pertaining to district, state, and national assessment;

You will learn about different types of instruments, several measurement concepts and issues, how to prepare your own multiple choice and performance assessments, and how to construct and evaluate scoring rubrics. You will also be become knowledgeable on a few of today's major assessment issues. You will acquire tools to help your students with notetaking, studying, and test taking. You will be able to talk with anyone about testing, secure in the knowledge that you have reviewed what prominent scholars in assessment think you should understand about a broad array of important topics.

This is a very hands-on, applied book. There are checklists, suggestions, guidelines, and very few formulas. We take the attitude that any means to gather information about students, whether objective or subjective, is an assessment. Thus, this book talks about teacher made tests, portfolios, and teacher notes in addition to standardized tests. We are the first to admit that this book has lots of breadth but not much depth. It is not intended to replace a semester long course or two on measurement. Rather it is designed to arm the busy teacher with some tools that will help with everyday survival in today's environment of high-stakes testing and assessment demands.

If you find this book helpful (or even if you don't), please take a look at the on-line journal *Practical Assessment Research and Evaluation* - http://ericae.net/pare. PARE's goal is to provide education professionals access to refereed articles that can have a positive

impact on assessment, research, evaluation, and teaching practice, especially at the local education agency (LEA) level. Many of the articles that appear in this book were originally published in PARE. We will continue to add more as they become available.

If you like any part of this book, please feel free to photocopy that portion or print it from the free on-line version and share it with your colleagues. Our goal is to get good, useful information into the hands of teachers.

Lawrence M. Rudner
William D. Schafer

Rudner, L. and W. Schafer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.
From the free on-line version. To order print copies call 800 229-4200

Table of Contents

Introduction[1]

Throughout this book, the term "test" is viewed as any of a variety of techniques that can capture what a person knows in response to a question. This includes standardized and large scale tests of achievement, teacher developed paper-and-pencil tests, classroom questions(including interactions to see whether students are ready to move on during instruction), performance tests - any system of collecting data where there is a “correct” response or responses that are better than others.

In this context, testing and teaching should be intertwined. The information provided by tests in their various forms, should be the tools that guide the instructional process, for both teacher and student. Twelve years ago, Rudman (1989) pointed out some instructional roles for educational tests. They haven’t changed:

Testing is a useful tool at the beginning of the school year.

It can help a teacher gain an overview of what students bring to new instruction. Test results early in the school year can help the teacher plan review material and identify potential issues to be faced. Examining past test results can help a teacher who is new to a specific school assess the school setting that he or she will work in as well as the expectations the school has for his or her students.

Testing can aid in decisions about grouping students in the class.

Testing can yield information that will aid the teacher in assigning specific students to instructional groups. The teacher can change the groups later after more teaching and testing has taken place.

Testing can be used to diagnose what individual pupils know.

No one source of data can be sufficient to assess what a pupil knows about school-related content. What is called for is a triangulation (corroboration) of several kinds of data drawn from various types of tests: standardized tests of achievement and aptitude, teacher-made quizzes, observations of behavior, informal interactions, and the like. Diagnosis does not necessarily mean prescription unless the data collected have demonstrated high reliability and validity, that is, you can trust them and they convey what you need to know in order to make instructional decisions about students.

Testing can help the teacher determine the pace of classroom instruction.

Teachers tend to use tests that they prepared themselves much more often than any other type of test to monitor what has been previously learned. These tests may take the form of oral questioning of the class or individual students, or paper-and-pencil tests. Systematic observations of a student applying a skill can be thought of as a form of performance testing. Tests used in these ways are prerequisites for determining how quickly

[1] Prepared by Lawrence M. Rudner and William D. Schafer

new material can be presented. These tests help the teacher gain a perspective of the range of attained learning as well as individual competence.

Tests can be used to help make promotion and retention decisions.

Many factors enter into the important decision of moving a student into the next grade. Intuition is an important part of any decision but that intuition is enhanced when coupled with data. Standardized tests, and records of classroom performance on less formal tests are essential for supplying much of the data upon which these decisions are based.

Test results are important devices to share information with boards of education, parents, and the general public through the media.

Classroom instruction depends upon a large support network. That network needs information if an adequate support level is to be maintained. Tests in various forms can supply that information. Informational needs vary among the support groups; specialized referrals for remediation and enrichment need test data for parental support and approval; effectiveness of educational planning is needed by boards of education: evidence which can be partially supplied by test data; financial support of existing programs by the general community needs evidence that can be supplied by test data.

Test results are useful tools for measuring the effectiveness of instruction and learning.

Various types of tests can be employed when measuring how effectively teaching impacts student learning. Learning when viewed in the aggregate can be viewed within a district at three levels; district, building, and classroom. Standardized tests are particularly useful at all three levels. These tests can be used in norm, criterion and objective-referenced modes. Tests written within the district for large-scale use can also supply information focused specifically on unique, local aspects of educational programs.

Hopefully, this book will give you the skills and knowledge needed to properly implement these instructional uses of test information. We have organized the chapters into three sections.

Section 1: Fundamental Concepts Common to All Assessments. This book starts with an examination of the fundamental concepts of testing. Several organizations and several projects have attempted to identify what teachers and administrators need to k now about testing. Here, we provide a synthesis of suggestions from these sources and present eleven basic non-statistical principles to provide a conceptual understanding of what tests can and cannot do. We then provide an overview of the basic measurement concepts of validity and reliability. A test is useless if the inferences based on the test are not reasonable, i.e. the test is not valid for the intended use. This chapter will help you judge the validity of an instrument. Similarly, a test is useless if the resultant scores contain a great deal of error, i.e. the test is not reliable. The next chapter discusses several ways to examine the reliability of a test. Test scores, by themselves, do not have any intrinsic meaning. We give meaning to scores by comparing them to scores of other children or by com paring the scores to

some criterion. This section ends includes a discussion of norm-referenced and criterion referenced tests.

This section also includes standardized and large scale assessments - typically the types of tests sponsored by state education agencies, reported in the popular press, and unfortunately, often inappropriately used as the sole measure to judge the worth of a school. We start with a discussion of the different types of scores used to report standardized test results. You will learn the advantages, disadvantages of each along with how the different types of scores should be used. A key feature of state assessments is that they are almost always accompanied by a careful delineation of endorsed educational goals. There should be no ambiguity with regard to what is covered by such tests. The next chapter discusses aligning one's instruction to the test and making the test into a valuable instructional planning tool. There is often a debate with regard to teaching to a test. Some argue that since the test identifies goals, teaching to the test is equivalent to teaching goals and should be done. Others argue that teaching to a test is an attempt to short circuit the educational process. The next chapter identifies a continuum of acceptable and unacceptable practices for preparing students to take standardized achievement tests. Lastly, with testing so prominent in the popular press, we provide an overview of some of the politics of national testing.

Section 2: Essential Concepts for Classroom Assessment. The most frequent and most valuable types of tests are those developed and used by classroom teachers. This section is designed to help you develop you write better multiple choice and better performance tests. You will learn to examine what it is that you want to assess, how to write questions that assess those concepts. Special attention is paid to the development of analytic and holistic scoring rubrics. Consistent with the view of testing as a form of data gathering and communication, chapters have been included on asking classroom questions as part of routine instruction and on writing comments on report cards.

Section 3: Essential Skills for Students. The last section is designed to help you help your students. Too often students appear to understand a concept in class, only to lose it the next day. We first provide some suggestions that you can implement that will help your students take better quality notes With better notes, students should be better organized and better prepared to meet academic expectations. Standardized tests are a reality. So is the fact that many students have misleading work patterns. We provide a chapter discussing common mistakes made by students and some teaching strategies that might help students overcome these mistakes. We end with a chapter actually written for students. It emphasizes the need for good study habits and it provides a few test-taking tips for different types of exams.

The Appendix includes two very import documents endorsed and developed by major organizations. The first, *Standards for Teacher Competence in Educational Assessment of Students* developed by the American Federation of Teachers, National Council on Measurement in Education, and the National Education Association, is intended to guide the preservice and inservice preparation of educators, the accreditation of preparation programs, and the future certification of all educators. We encourage you to compare your skills and knowledge against these standards. The second documents *Rights and Responsibilities of Test Takers: Guidelines and Expectations* contains the best judgments of testing professionals about

the reasonable expectations that those involved in the testing enterprise - test producers, test users, and test takers - should have of each other. The document is applicable to classroom tests as well as standardized tests.

Reference

Rudman, Herbert C. (1989). Integrating Testing with Teaching. *Practical Assessment, Research & Evaluation*, 1(6). Available online: http://ericae.net/pare/getvn.asp?v=1&n=6.

Fundamental Concepts Common to All Assessments

Fundamental Assessment Principles for Teachers and School Administrators[2]

While several authors have argued that there are a number of "essential" assessment concepts, principles, techniques, and procedures that teachers and administrators need to know about (e.g. Calfee & Masuda,1997; Cizek, 1997; Ebel, 1962; Farr & Griffin, 1973; Fleming & Chambers, 1983; Gullickson, 1985, 1986; Mayo, 1967; McMillan, 2001; Sanders & Vogel, 1993; Schafer, 1991; Stiggins & Conklin, 1992), there continues to be relatively little emphasis on assessment in the preparation of, or professional development of, teachers and administrators (Stiggins, 2000). In addition to the admonitions of many authors, there are established professional standards for assessment skills of teachers (*Standards for Teacher Competence in Educational Assessment of Students* (1990), a framework of assessment tasks for administrators (Impara & Plake, 1996), the Code of Professional Responsibilities in Educational Measurement (1995), the Code of Fair Testing Practices (1988), and the new edition *of Standards for Educational and Psychological Testing* (1999). If that isn''t enough information, a project directed by Arlen Gullickson at The Evaluation Center of Western Michigan University will publish standards for evaluations of students in the near future.

The purpose of this chapter is to use suggestions and guidelines from these sources, in light of current assessment demands and contemporary theories of learning and motivation, to present eleven "basic principles" to guide the assessment training and professional development of teachers and administrators. That is, what is it about assessment, whether large-scale or classroom, that is fundamental for effective understanding and application? What are the "big ideas" that, when well understood and applied, will effectively guide good assessment practices, regardless of the grade level, subject matter, developer, or user of the results? As Jerome Bruner stated it many years ago in his classic, *The Process of Education*: "……the curriculum of a subject should be determined by the most fundamental understanding that can be achieved of the underlying principles that give structure to that subject." (Bruner, 1960, p.31). What principles, in other words, provide the most essential, fundamental "structure" of assessment knowledge and skills that result in effective educational practices and improved student learning?

ASSESSMENT IS INHERENTLY A PROCESS OF PROFESSIONAL JUDGMENT.

The first principle is that professional judgment is the foundation for assessment and, as such, is needed to properly understand and use all aspects of assessment. The measurement of student performance may seem "objective" with such practices as machine scoring and multiple-choice test items, but even these approaches are based on professional assumptions and values. Whether that judgment occurs in constructing test questions, scoring essays**,** creating rubrics, grading participation, combining scores, or interpreting standardized test scores, the essence of the process is making professional interpretations

---

[2] Written by James H. McMillan

Rudner, L. and W. Schafer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.
From the free on-line version. To order print copies call 800 229-4200

and decisions. Understanding this principle helps teachers and administrators realize the importance of their own judgments and those of others in evaluating the quality of assessment and the meaning of the results.

ASSESSMENT IS BASED ON SEPARATE BUT RELATED PRINCIPLES OF MEASUREMENT EVIDENCE AND EVALUATION.

It is important to understand the difference between measurement evidence (differentiating degrees of a trait by description or by assigning scores) and evaluation (interpretation of the description or scores). Essential measurement evidence skills include the ability to understand and interpret the meaning of descriptive statistical procedures, including variability, correlation, percentiles, standard scores, growth-scale scores, norming, and principles of combining scores for grading. A conceptual understanding of these techniques is needed (not necessarily knowing how to compute statistics) for such tasks as interpreting student strengths and weaknesses, reliability and validity evidence, grade determination, and making admissions decisions. Schafer (1991) has indicated that these concepts and techniques comprise part of an essential language for educators. They also provide a common basis for communication about "results," interpretation of evidence, and appropriate use of data. This is increasingly important given the pervasiveness of standards-based, high-stakes, large-scale assessments. Evaluation concerns merit and worth of the data as applied to a specific use or context. It involves what Shepard (2000) has described as the systematic analysis of evidence. Like students, teachers and administrators need analysis skills to effectively interpret evidence and make value judgments about the meaning of the results.

ASSESSMENT DECISION-MAKING IS INFLUENCED BY A SERIES OF TENSIONS.

Competing purposes, uses, and pressures result in tension for teachers and administrators as they make assessment-related decisions. For example, good teaching is characterized by assessments that motivate and engage students in ways that are consistent with their philosophies of teaching and learning and with theories of development, learning and motivation. Most teachers want to use constructed-response assessments because they believe this kind of testing is best to ascertain student understanding. On the other hand, factors external to the classroom, such as mandated large-scale testing, promote different assessment strategies, such as using selected-response tests and providing practice in objective test-taking (McMillan & Nash, 2000). Further examples of tensions include the following.

- Learning vs auditing
- Formative (informal and ongoing) vs summative (formal and at the end)
- Criterion-referenced vs norm-referenced
- Value-added vs. absolute standards
- Traditional vs alternative
- Authentic vs contrived
- Speeded tests vs power tests
- Standardized tests vs classroom tests

These tensions suggest that decisions about assessment are best made with a full understanding of how different factors influence the nature of the assessment. Once all the alternatives understood, priorities need to be made; trade-offs are inevitable. With an appreciation of the tensions teachers and administrators will hopefully make better informed, better justified assessment decisions.

ASSESSMENT INFLUENCES STUDENT MOTIVATION AND LEARNING.

Grant Wiggins (1998) has used the term 'educative assessment' to describe techniques and issues that educators should consider when they design and use assessments. His message is that the nature of assessment influences what is learned and the degree of meaningful engagement by students in the learning process. While Wiggins contends that assessments should be authentic, with feedback and opportunities for revision to improve rather than simply audit learning, the more general principle is understanding how different assessments affect students. Will students be more engaged if assessment tasks are problem-based? How do students study when they know the test consists of multiple-choice items? What is the nature of feedback, and when is it given to students? How does assessment affect student effort? Answers to such questions help teachers and administrators understand that assessment has powerful effects on motivation and learning. For example, recent research summarized by Black & Wiliam (1998) shows that student self-assessment skills, learned and applied as part of formative assessment, enhances student achievement.

ASSESSMENT CONTAINS ERROR.

Teachers and administrators need to not only know that there is error in all classroom and standardized assessments, but also more specifically how reliability is determined and how much error is likely. With so much emphasis today on high-stakes testing for promotion, graduation, teacher and administrator accountability, and school accreditation, it is critical that all educators understand concepts like standard error of measurement, reliability coefficients, confidence intervals, and standard setting. Two reliability principles deserve special attention. The first is that reliability refers to scores, not instruments. Second, teachers and administrators need to understand that, typically, error is underestimated. A recent paper by Rogosa (1999), effectively illustrates the concept of underestimation of error by showing in terms of percentile rank probable true score hit-rate and test-retest results.

GOOD ASSESSMENT ENHANCES INSTRUCTION.

Just as assessment impacts student learning and motivation, it also influences the nature of instruction in the classroom. There has been considerable recent literature that has promoted assessment as something that is integrated with instruction, and not an activity that merely audits learning (Shepard, 2000). When assessment is integrated with instruction it informs teachers about what activities and assignments will be most useful, what level of teaching is most appropriate, and how summative assessments provide diagnostic information. For instance, during instruction activities informal, formative assessment helps teachers know when to move on, when to ask more questions, when to give more examples, and what responses to student questions are most appropriate. Standardized test scores,

when used appropriately, help teachers understand student strengths and weaknesses to target further instruction.

GOOD ASSESSMENT IS VALID.

Validity is a concept that needs to be fully understood. Like reliability, there are technical terms and issues associated with validity that are essential in helping teachers and administrators make reasonable and appropriate inferences from assessment results (e.g., types of validity evidence, validity generalization, construct underrepresentation, construct-irrelevant variance, and discriminant and convergent evidence). Of critical importance is the concept of evidence based on consequences, a new major validity category in the recently revised *Standards*. Both intended and unintended consequences of assessment need to be examined with appropriate evidence that supports particular arguments or points of view. Of equal importance is getting teachers and administrators to understand their role in gathering and interpreting validity evidence.

GOOD ASSESSMENT IS FAIR AND ETHICAL.

Arguably, the most important change in the recently published *Standards* is an entire new major section entitled "Fairness in Testing." The *Standards* presents four views of fairness: as absence of bias (e.g., offensiveness and unfair penalization), as equitable treatment, as equality in outcomes, and as opportunity to learn. It includes entire chapters on the rights and responsibilities of test takers, testing individuals of diverse linguistic backgrounds, and testing individuals with disabilities or special needs. Three additional areas are also important:

- Student knowledge of learning targets and the nature of the assessments prior to instruction (e.g., knowing what will be tested, how it will be graded, scoring criteria, anchors, exemplars, and examples of performance).
- Student prerequisite knowledge and skills, including test-taking skills.
- Avoiding stereotypes.

GOOD ASSESSMENTS USE MULTIPLE METHODS.

Assessment that is fair, leading to valid inferences with a minimum of error, is a series of measures that show student understanding through multiple methods. A complete picture of what students understand and can do is put together in pieces comprised by different approaches to assessment. While testing experts and testing companies stress that important decisions should not be made on the basis of a single test score, some educators at the local level, and some (many?) politicians at the state at the national level, seem determined to violate this principle. There is a need to understand the entire range of assessment techniques and methods, with the realization that each has limitations.

GOOD ASSESSMENT IS EFFICIENT AND FEASIBLE.

Teachers and school administrators have limited time and resources. Consideration must be given to the efficiency of different approaches to assessment, balancing needs to implement methods required to provide a full understanding with the time needed to develop and implement the methods, and score results. Teacher skills and knowledge are important to consider, as well as the level of support and resources.

GOOD ASSESSMENT APPROPRIATELY INCORPORATES TECHNOLOGY.

As technology advances and teachers become more proficient in the use of technology, there will be increased opportunities for teachers and administrators to use computer-based techniques (e.g., item banks, electronic grading, computer-adapted testing, computer-based simulations), Internet resources, and more complex, detailed ways of reporting results. There is, however, a danger that technology will contribute to the mindless use of new resources, such as using items on-line developed by some companies without adequate evidence of reliability, validity, and fairness, and crunching numbers with software programs without sufficient thought about weighting, error, and averaging.

To summarize, what is most essential about assessment is understanding how general, fundamental assessment principles and ideas can be used to enhance student learning and teacher effectiveness. This will be achieved as teachers and administrators learn about conceptual and technical assessment concepts, methods, and procedures, for both large-scale and classroom assessments, and apply these fundamentals to instruction.

Notes:

Earlier versions of this paper were presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 24, 2000 and published as McMillan, James H. (2000). Fundamental Assessment Principles for Teachers and School Administrators. *Practical Assessment, Research & Evaluation*, 7(8). Available online: http://ericae.net/pare/getvn.asp?v=7&n=8.

References

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan,* 80(2), 139-148.

Bruner, J. S. (1960). *The process of education.* NY: Vintage Books.

Calfee, R. C., & Masuda, W. V. (1997). Classroom assessment as inquiry. In G. D. Phye (Ed.) *Handbook of classroom assessment: Learning, adjustment, and achievement.* NY: Academic Press.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.) *Handbook of classroom assessment: Learning, adjustment, and achievement.* NY: Academic Press.

Code of fair testing practices in education (1988). Washington, DC: Joint Committee on Testing Practices (American Psychological Association). Available http://ericae.net/code.htm

Code of professional responsibilities in educational measurement (1995). Washington, DC: National Council on Measurement in Education. Available http://www.unl.edu/buros/article2.html

Ebel, R. L. (1962). Measurement and the teacher. *Educational Leadership*, 20, 20-24.

Farr, R., & Griffin, M. (1973). Measurement gaps in teacher education. *Journal of Research and Development in Education*, 7(1), 19-28.
Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway (Ed.), *Testing in the schools*, San Francisco: Jossey-Bass.
Gullickson, A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research, 79*(2), 96-100.
Gullickson, A. R. (1996). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement, 23*(4), 347-354.
Impara, J. C., & Plake, B. S. (1996). Professional development in student assessment for educational administrators. *Educational Measurement: Issues and Practice, 15*(2), 14-19.
Mayo, S. T. (1967). Pre-service preparation of teachers in educational measurement. U.S. Department of Health, Education and Welfare. Washington, DC: Office of Education/Bureau of Research.
McMillan, J. H. (2001). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Publishing Company.
McMillan, J. H., & Nash, S. (2000). Teachers' classroom assessment and grading decision making. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.
Rogosa, D. (1999). How accurate are the STAR national percentile rank scores for individual students? - An interpretive guide. Palo Alto, CA: Stanford University.
Sanders, J. R., & Vogel, S. R. (1993). The development of standards for teacher competence in educational assessment of students, in S. L. Wise (Ed.), *Teacher training in measurement and assessment skills*, Lincoln, NB: Burros Institute of Mental Measurements.
Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice*, 10, (1), 3-6.
Shepard, L. A. (2000). The role of assessment in a learning culture. Paper presented at the Annual Meeting of the American Educational Research Association. Available http://www.aera.net/meeting/am2000/wrap/praddr01.htm
Standards for educational and psychological testing (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on 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. Available: http://www.unl.edu/buros/article3.html
Stiggins, R. J. (2000). Classroom assessment: A history of neglect, a future of immense potential. Paper presented at the Annual Meeting of the American Educational Research Association.
Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany, NY: State University of New York Press, Albany.
Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

Traditional and Modern Concepts of Validity[1]

Test validity refers to the degree with which the *inferences* based on test scores are meaningful, useful, and appropriate. Thus test validity is a characteristic of a test when it is administered to a particular population. Validating a test refers to accumulating empirical data and logical arguments to show that the inferences are indeed appropriate.

This chapter introduces the modern concepts of validity advanced by the late Samuel Messick (1989, 1996a, 1996b). We start with a briefly review the traditional methods of gathering validity evidence.

TRADITIONAL CONCEPT OF VALIDITY

Traditionally, the various means of accumulating validity evidence have been grouped into three categories -- content-related, criterion-related, and construct-related evidence of validity. These broad categories are a convenient way to organize and discuss validity evidence. There are no rigorous distinctions between them; they are not distinct types of validity. Evidence normally identified with the criterion-related or content-related categories, for example, may also be relevant in the construct-related evidence

Criterion-related validity evidence - seeks to demonstrate that test scores are systematically related to one or more outcome criteria. In terms of an achievement test, for example, criterion-related validity may refer to the extent to which a test can be used to draw inferences regarding achievement. Empirical evidence in support of criterion-related validity may include a comparison of performance on the test against performance on outside criteria such as grades, class rank, other tests and teacher ratings.

Content-related validity evidence - refers to the extent to which the test questions represent the skills in the specified subject area. Content validity is often evaluated by examining the plan and procedures used in test construction. Did the test development procedure follow a rational approach that ensures appropriate content? Did the process ensure that the collection of items would represent appropriate skills?

Construct-related validity evidence - refers to the extent to which the test measures the "right" psychological constructs. Intelligence, self-esteem and creativity are examples of such psychological traits. Evidence in support of construct-related validity can take many forms. One approach is to demonstrate that the items within a measure are inter-related and therefore measure a single construct. Inter-item correlation and factor analysis are often used to demonstrate relationships among the items. Another approach is to demonstrate that the test behaves as one would expect a measure of the construct to behave. For example, one might expect a measure of creativity to show a greater correlation with a measure of artistic ability than with a measure of scholastic achievement.

---

[1] Written by Amy Brualdi

MODERN CONCEPT OF VALIDITY

Messick (1989, 1996a) argues that the traditional conception of validity is fragmented and incomplete especially because it fails to take into account both evidence of the value implications of score meaning as a basis for action and the social consequences of score use. His modern approach views validity as a unified concept which places a heavier emphasis on how a test is used. Six distinguishable aspects of validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. In effect, these six aspects conjointly function as general validity criteria or standards for all educational and psychological measurement. These six aspects must be viewed as interdependent and complementary forms of validity evidence and not viewed as separate and substitutable validity types.

Content A key issue for the content aspect of validity is determining the knowledge, skills, and other attributes to be revealed by the assessment tasks. Content standards themselves should be relevant and representative of the construct domain. Increasing achievement levels or performance standards should reflect increases in complexity of the construct under scrutiny and not increasing sources of construct-irrelevant difficulty (Messick, 1996a).

Substansive The substansive aspect of validity emphasizes the verification of the domain processes to be revealed in assessment tasks. These can be identified through the use of substansive theories and process modeling (Embretson, 1983; Messick 1989). When determining the substansiveness of test, one should consider two points. First, the assessment tasks must have the ability to provide an appropriate sampling of domain processes in addition to traditional coverage of domain content. Also, the engagement of these sampled in these assessment tasks must be confirmed by the accumulation of empirical evidence.

Structure Scoring models should be rationally consistent with what is known about the structural relations inherent in behavioral manifestations of the construct in question (Loevinger, 1957). The manner in which the execution of tasks are assessed and scored should be based on how the implicit processes of the respondent's actions combine dynamically to produce effects. Thus, the internal structure of the assessment should be consistent with what is known about the internal structure of the construct domain (Messick, 1989).

Generalizability Assessments should provide representative coverage of the content and processes of the construct domain. This allows score interpretations to be broadly generalizable within the specified construct. Evidence of such generalizability depends on the tasks' degree of correlation with other tasks that also represent the construct or aspects of the construct.

External Factors The external aspects of validity refers to the extent that the assessment scores' relationship with other measures and nonassessment behaviors reflect the expected high, low, and interactive relations implicit in the specified construct. Thus, the score interpretation is substantiated externally by appraising the degree to which empirical relationships are consistent with that meaning.

Consequential Aspects of Validity It is important to accrue evidence of such positive consequences as well as evidence that adverse consequences are minimal. The consequential aspect of validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use. This type of investigation is especially important when it concerns adverse consequences for individuals and groups that are associated with bias in scoring and interpretation.

These six aspects of validity apply to all educational and psychological measurement; most score-based interpretations and action inferences either invoke these properties or assume them, explicitly or tacitly. The challenge in test validation, then, is to link these inferences to convergent evidence which support them as well as to discriminant evidence that discount plausible rival inferences.

SOURCES OF INVALIDITY

Two major threats to test validity are worth noting, especially with today's emphasis on high-stakes performance tests.

Construct underrepresentation indicates that the tasks which are measured in the assessment fail to include important dimensions or facets of the construct. Therefore, the test results are unlikely to reveal a student's true abilities within the construct which the test was indicated as having been measured.

Construct-irrelevant variance means that the test measures too many variables, many of which are irrelevant to the interpreted construct. This type of invalidity can take two forms, "construct-irrelevant easiness" and "construct-irrelevant difficulty." "Construct-irrelevant easiness" occurs when extraneous clues in item or task formats permit some individuals to respond correctly or appropriately in ways that are irrelevant to the construct being assessed; "construct-irrelevant difficulty" occurs when extraneous aspects of the task make the task irrelevantly difficult for some individuals or groups. While the first type of construct irrelevant variance causes one to score higher than one would under normal circumstances, the latter causes a notably lower score.

Because there is a relative dependence of task responses on the processes, strategies, and knowledge that are implicated in task performance, one should be able to identify through cognitive-process analysis the theoretical mechanisms underlying task performance (Embretson, 1983).

REFERENCES AND RECOMMENDED READING

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.

Embretson (Whitely), S. Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Fredericksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694 (Monograph Supplement 9).

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1996a). Standards-based score interpretation: Establishing valid grounds for valid inferences. *Proceedings of the joint conference on standard setting for large scale assessments*, Sponsored by National Assessment Governing Board and The National Center for Education Statistics. Washington, DC: Government Printing Office.

Messick, S. (1996b). Validity of Performance Assessment. In Philips, G. (1996). *Technical Issues in Large-Scale Performance Assessment*. Washington, DC: National Center for Educational Statistics.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229-258.

Reliability[1]

All tests contain error. This is true for tests in both the physical sciences and psychological tests. In measuring length with a ruler, for example, there may be systematic error associated with where the zero point is printed on the ruler and random error associated with your eye's ability to read the marking and extrapolate between the markings. It is also possible that the length of the object can vary over time and environment (e.g., with changes in temperature). One goal in assessment is to keep these errors down to levels that are appropriate for the purposes of the test. High-stakes tests, such as licensure examinations, need to have very little error. Classroom tests can tolerate more error as it is fairly easy to spot and correct mistakes made during the testing process. Reliability focuses only on the degree of errors that are nonsystematic, called random errors.

Reliability has been defined in different ways by different authors. Perhaps the best way to look at reliability is the extent to which the measurements resulting from a test are the result of characteristics of those being measured. For example, reliability has elsewhere been defined as "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" (Berkowitz, Wolkowitz, Fitch, and Kopriva, 2000). This definition will be satisfied if the scores are indicative of properties of the test takers; otherwise they will vary unsystematically and not be repeatable or dependable.

Reliability can also be viewed as an indicator of the absence of random error when the test is administered. When random error is minimal, scores can be expected to be more consistent from administration to administration.

Technically, the theoretical definition of reliability is the proportion of score variance that is caused by systematic variation in the population of test-takers. This definition is population-specific. If there is greater systematic variation in one population than another, such as in all public school students compared with only eighth-graders, the test will have greater reliability for the more varied population. This is a consequence of how reliability is defined. Reliability is a joint characteristic of a test and examinee group, not just a characteristic of a test. Indeed, reliability of any one test varies from group to group. Therefore, the better research studies will report the reliability for their sample as well as the reliability for noming groups as presented by the test publisher.

This chapter discusses sources of error, several approaches toward estimating reliability, and several ways to make your tests more reliable.

SOURCES OF ERROR

[1] Written by Lawrence M. Rudner and William D. Schafer

Rudner, Lawrence M. & Schafer, William D. (2001). *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.

There are three major sources of error: factors in the test itself, factors in the students taking the test, and scoring factors.

Most tests contain a collection of items that represent particular skills. We typically generalize from each item to all items like that item. For example, if a student can solve several problems like 7 times 8 then we may generalize his or her ability to multiply single-digit integers. We also generalize from the collection of items to a broader domain. If a student does well on a test of addition, subtraction, multiplication and division of fractions, then we may generalize and conclude that the student is able to perform fraction operations. But error may be introduced by the selection of particular items to represent the skills and domains. The particular cross section of test content that is included in the specific items on the test will vary with each test form, introducing sampling error and limiting the dependability of the test, since we are generalizing to unobserved data, namely, ability across all items that could have been on the test. On basic arithmetic skills, one would expect the content to be fairly similar and thus building a highly reliable test is relatively easy. As the skills and domains become more complex, more errors are likely introduced by sampling of items. Other sources of test error include the effectiveness of the distractors (wrong options) in multiple choice tests, partially correct distractors, multiple correct answers, and difficulty of the items relative to the student's ability.

As human beings, students are not always consistent and also introduce error into the testing process. Whether a test is intended to measure typical or optimal student performance, changes in such things as student's attitudes, health, and sleep may affect the quality of their efforts and thus their test taking consistency. For example, test takers may make careless errors, misinterpret test instructions, forget test instructions, inadvertently omit test sections, or misread test items.

Scoring errors are a third potential source of error. On objective tests, the scoring is mechanical and scoring error should be minimal. On constructed response items, sources of error include clarity of the scoring rubrics, clarity of what is expected of the student, and a host of rater errors. Raters are not always consistent, sometimes change their criteria while scoring, and are subject to biases such as the halo effect, stereotyping, perception differences, leniency/stringency error, and scale shrinkage (see Rudner, 1992).

MEASURES OF RELIABILITY

It is impossible to calculate a reliability coefficient that conforms to the theoretical definition. Recall, the theoretical definition depends on knowing the degree to which a population of examinees vary in their true achievement (or whatever the test measures). But if we knew that, then we wouldn't need the test! Instead, there are several statistics (coefficients) commonly used to estimate the stability of a set of test scores for a group of examinees: test-retest, split-half reliability, alternate form reliability, and measures of internal consistency are the most common.

Reliability is a joint characteristic of a test and examinee group

Test-retest reliability. A test-retest reliability coefficient is obtained by administering the same test twice and correlating the scores. In concept, it is an excellent measure of score consistency. One is directly measuring consistency from administration to administration. This coefficient is not recommended in practice, however, because of its problems and limitations. It requires two administrations of the same test with the same group of individuals. This is expensive and not a good use of people's time. If the time interval is short, people may be overly consistent because they remember some of the question and their responses. If the interval is long, then the results are confounded with learning and maturation, that is, changes in the persons, themselves.

Split-half reliability. As the name suggests, split-half reliability is a coefficient obtained by dividing a test into halves, correlating the scores on each half, and then correcting for length (longer tests tend to be more reliable). The split can be based on odd versus even numbered items, randomly selecting items, or manually balancing content and difficulty. This approach has an advantage in that it only requires a single test administration. Its weakness is that the resultant coefficient will vary as a function of how the test was split. It is also not appropriate on tests where speed is a factor (that is, where students' scores are influenced by how many items they reached in the allotted time).

Internal consistency. Internal consistency focuses on the degree to which the individual items are correlated with each other and is thus often called homogeneity. Several statistics fall within this category. The best known are Cronbach's alpha, the Kuder-Richardson Formula 20 (KR-20) and the Kuder-Richardson Formula 21 (KR-21). Most testing programs that report data from one administration of a test to students do so using Cronbach's alpha, which is functionally equivalent to KR-20.

The advantages of these statistics are that they only require one test administration and that they do not depend on a particular split of items. The disadvantage is that they are most applicable when the test measures a single skill area.

Requiring only the test mean, standard deviation (or variance), and the number of items, the Kuder-Richardson formula 21 is an extremely simple reliability formula. While it will almost always provide coefficients that are lower than KR-20, its simplicity makes it's a very useful estimate of reliability, especially for evaluating some classroom-developed tests. However, it should not be used if the test has items that are scored other than just zero or one.

$$KR\,21 = \frac{k}{k-1}\left(1 - \frac{M(k-M)}{k\sigma^2}\right)$$

Where M is the mean, k is the number of items, and σ^2 is the test variance.

Alternate-form reliability. Most standardized tests provide equivalent forms that can be used interchangeably. These alternative forms are typically matched in terms of content and difficulty. The correlation of scores on pairs of alternative forms for the same examinees provides another measure of consistency or reliability. Even with the best test and item specifications, each test would contain slightly different content and, as with test-

retest reliability, maturation and learning may confound the results. However, the use of different items in the two forms conforms to our goal of including the extent to which item sets contribute to random errors in estimating test reliability.

How High Should Reliability Be?

Most large-scale tests report reliability coefficients that exceed .80 and often exceed .90. The questions to ask are 1) what are the consequences of the test and 2) is the group used to compute the reported reliability like my group.

If the consequences are high, as in tests used for special education placement, high school graduation and certification, then the internal consistency reliability needs to be quite high - at least above .90, preferably above.95. Misclassifications due to measurement error should be kept to a minimum. And please note that no test should ever be used by itself to make an important decision for anyone.

Classroom tests seldom need to have exceptionally high reliability coefficients. As more students master the content, test variability will go down and so will the coefficients from internal measures of reliability. Further, classroom tests don't need exceptionally high reliability coefficients. As teachers, you see the child all day and have gathered input from a variety of information sources. Your knowledge and judgment, used along with information from the test, provides superior information. If a test is not reliable or it is not accurate for an individual, you can and should make the appropriate corrections. A reliability coefficient of .50 or .60 may suffice.

Again, reliability is a joint characteristic of a test and examinee group, not just a characteristic of a test. Thus, reliability also needs to be evaluated in terms of the examinee group. A test with a reliability of .92 when administered to students in 4th, 5th, and 6th grades will not have as high a reliability when administered just to a group of 4th graders.

IMPROVING TEST RELIABILITY

Developing better tests with less random measurement error is better than simply documenting the amount of error. Measurement error is reduced by writing items clearly, making the instructions easily understood, adhering to proper test administration, and consistent scoring. Because a test is a sample of the desired skills and behaviors, longer tests, which are larger samples, will be more reliable. A one-hour end-of-unit exam will be more reliable than a 5 minute pop-quiz. (Note that pop quizzes should be discouraged. By using them, a teacher is not only using assessments punitively, but is also missing the opportunity to capitalize on student preparation as an instructional activity.)

A COMMENT ON SCORING

What do you do if a child makes careless mistakes on a test? On one hand, you want your students to learn to follow directions, to think through their work, to check their work, and to be careful. On the other hand, tests are supposed to reflect what a student knows. Further, a low score due to careless mistakes is not the same as a low score due to lack of knowledge.

Especially in the elementary grades, a miserable test due to careless mistakes should not dramatically lower a student's grade for the semester. The semester grade should reflect what the student has achieved, since that is the meaning it will convey to others. We advocate keeping two sets of records, especially in the elementary grades. One set reflects production, and the other reflecting achievement. The teacher then has the needed data to apply good judgment in conferencing with parents and for determining semester grades.

References and Recommended Reading

Anastasi, A. (1988). *Psychological Testing*. New York, New York: MacMillan Publishing Company.

Berkowitz, David, Barbara Wolkowitz, Rebecca Fitch, and Rebecca Kopriva (2000). *The Use of Tests as Part of High-Stakes Decision-Making for Students: A Resource Guide for Educators and Policy-Makers*. Washington, DC: U.S. Department of Education.

Lyman, Howard B. (1993). *Test Scores and What They Mean*. Boston: Allyn and Bacon.

McMillan, James H. (2001). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Publishing Company.

Nunnally, Jum, C. (1967). *Psychometric Theory*. New York: McGraw-Hill Book Company, Chapter 6 and 7.

Popham, W. James (1998). *Classroom Assessment, What Teachers Need to Know*. Boston, Allyn and Bacon.

Rudner, Lawrence M. (1992). Reducing Errors Due to the Use of Judges. *Practical Assessment, Research & Evaluation*, 3(3). Available online: http://ericae.net/pare/getvn.asp?v=3&n=3.

Norm- and Criterion-Referenced Testing[1]

Tests can be categorized into two major groups: norm-referenced tests and criterion-referenced tests. These two tests differ in their intended purposes, the way in which content is selected, and the scoring process which defines how the test results must be interpreted. This brief paper will describe the differences between these two types of assessments and explain the most appropriate uses of each.

INTENDED PURPOSES

The major reason for using a norm-referenced tests (NRT) is to classify students. NRTs are designed to highlight achievement differences between and among students to produce a dependable rank order of students across a continuum of achievement from high achievers to low achievers (Stiggins, 1994). School systems might want to classify students in this way so that they can be properly placed in remedial or gifted programs. These types of tests are also used to help teachers select students for different ability level reading or mathematics instructional groups.

With norm-referenced tests, a representative group of students is given the test prior to its availability to the public. The scores of the students who take the test after publication are then compared to those of the norm group. Tests such as the California Achievement Test (CTB/McGraw-Hill), the Iowa Test of Basic Skills (Riverside), and the Metropolitan Achievement Test (Psychological Corporation) are normed using a national sample of students. Because norming a test is such an elaborate and expensive process, the norms are typically used by test publishers for 7 years. All students who take the test during that seven year period have their scores compared to the original norm group.

While norm-referenced tests ascertains the rank of students, criterion-referenced tests (CRTs) determine "...what test takers can do and what they know, not how they compare to others (Anastasi, 1988, p. 102). CRTs report how well students are doing relative to a pre-determined performance level on a specified set of educational goals or outcomes included in the school, district, or state curriculum.

Educators or policy makers may choose to use a CRT when they wish to see how well students have learned the knowledge and skills which they are expected to have mastered. This information may be used as one piece of information to determine how well the student is learning the desired curriculum and how well the school is teaching that curriculum.

Both NRTs and CRTs can be standardized. The U.S. Congress, Office of Technology Assessment (1992) defines a standardized test as one that uses uniform procedures for administration and scoring in order to assure that the results from different people are comparable. Any kind of test--from multiple choice to essays to oral examinations--can be standardized if uniform scoring and administration are used (p. 165). This means that the comparison of student scores is possible. Thus, it can be assumed that two students who

---

[1] Written by Linda A. Bond

receive the identical scores on the same standardized test demonstrate corresponding levels of performance. Most national, state and district tests are standardized so that every score can be interpreted in a uniform manner for all students and schools.

SELECTION OF TEST CONTENT

Test content is an important factor choosing between an NRT test and a CRT test. The content of an NRT test is selected according to how well it ranks students from high achievers to low. The content of a CRT test is determined by how well it matches the learning outcomes deemed most important. Although no test can measure everything of importance, the content selected for the CRT is selected on the basis of its significance in the curriculum while that of the NRT is chosen by how well it discriminates among students.

Any national, state or district test communicates to the public the skills that students should have acquired as well as the levels of student performance that are considered satisfactory. Therefore, education officials at any level should carefully consider content of the test which is selected or developed. Because of the importance placed upon high scores, the content of a standardized test can be very influential in the development of a school's curriculum and standards of excellence.

NRTs have come under attack recently because they traditionally have purportedly focused on low level, basic skills. This emphasis is in direct contrast to the recommendations made by the latest research on teaching and learning which calls for educators to stress the acquisition of conceptual understanding as well as the application of skills. The National Council of Teachers of Mathematics (NCTM) has been particularly vocal about this concern. In an NCTM publication (1991), Romberg (1989) cited that "a recent study of the six most commonly used commercial achievement tests found that at grade 8, on average, only 1 percent of the items were problem solving while 77 percent were computation or estimation" (p. 8).

In order to best prepare their students for the standardized achievement tests, teachers usually devote much time to teaching the information which is found on the standardized tests. This is particularly true if the standardized tests are also used to measure an educator's teaching ability. The result of this pressure placed upon teachers for their students to perform well on these tests has resulted in an emphasis on low level skills in the classroom (Corbett & Wilson, 1991). With curriculum specialists and educational policy makers alike calling for more attention to higher level skills, these tests may be driving classroom practice in the opposite direction of educational reform.

TEST INTERPRETATION

As mentioned earlier, a student's performance on an NRT is interpreted in relation to the performance of a large group of similar students who took the test when it was first normed. For example, if a student receives a percentile rank score on the total test of 34, this means that he or she performed as well or better than 34% of the students in the norm group. This type of information can useful for deciding whether or not students need remedial assistance or is a candidate for a gifted program. However, the score gives little

information about what the student actually knows or can do. The validity of the score in these decision processes depends on whether or not the content of the NRT matches the knowledge and skills expected of the students in that particular school system.

It is easier to ensure the match to expected skills with a CRT. CRTs give detailed information about how well a student has performed on each of the educational goals or outcomes included on that test. For instance, "... a CRT score might describe which arithmetic operations a student can perform or the level of reading difficulty he or she can comprehend" (U.S. Congress, OTA, 1992, p. 170). As long as the content of the test matches the content that is considered important to learn, the CRT gives the student, the teacher, and the parent more information about how much of the valued content has been learned than an NRT.

SUMMARY

Public demands for accountability, and consequently for high standardized tests scores, are not going to disappear. In 1994, thirty-one states administered NRTs, while thirty-three states administered CRTs. Among these states, twenty-two administered both. Only two states rely on NRTs exclusively, while one state relies exclusively on a CRT. Acknowledging the recommendations for educational reform and the popularity of standardized tests, some states are designing tests that "reflect, insofar as possible, what we believe to be appropriate educational practice" (NCTM, 1991, p.9). In addition to this, most states also administer other forms of assessment such as a writing sample, some form of open-ended performance assessment or a portfolio (CCSSO/NCREL, 1994).

Before a state can choose what type of standardized test to use, the state education officials will have to consider if that test meets three standards. These criteria are whether the assessment strategy(ies) of a particular test matches the state's educational goals, addresses the content the state wishes to assess, and allows the kinds of interpretations state education officials wish to make about student performance. Once they have determined these three things, the task of choosing between the NRT and CRT will becomes easier.

REFERENCES

Anastasi, A. (1988). *Psychological Testing.* New York, New York: MacMillan Publishing Company.

Corbett, H.D. & Wilson, B.L. (1991). *Testing, Reform and Rebellion.* Norwood, New Jersey: Ablex Publishing Company.

Romberg, T.A., Wilson, L. & Mamphono Khaketla (1991). "The Alignment of Six Standardized Tests with NCTM Standards", an unpublished paper, University of Wisconsin-Madison. In Jean Kerr Stenmark (ed; 1991). *Mathematics Assessment: Myths, Models, Good Questions, and Practical Suggestions.* The National Council of Teachers of Mathematics (NCTM)

Stenmark, J.K (ed; 1991). *Mathematics Assessment: Myths, Models, Good Questions, and Practical Suggestions.* Edited by. Reston, Virginia: The National Council of Teachers of Mathematics (NCTM)

Stiggins, R.J. (1994). Student-Centered Classroom Assessment. New York: Merrill

U.S. Congress, Office of Technology Assessment (1992). *Testing in America's Schools: Asking the Right Questions.* OTA-SET-519 (Washington, D.C.: U.S. Government Printing Office)

Table 1. Appropriate Uses of Norm-referenced and Criterion-referenced Tests [1]

Purpose	Test	Examples	Primary users
To compare achievement of local students to achievement of students in the nation, state, or other districts in a given year.	NRT	A comparison of achievement of local schools' 3rd graders to achievement of 3rd graders throughout the nation.	Central office, (including school boards), parents
To compare achievement of subgroups of local students to achievement of similar subgroups in the nation, state, or other districts in a given year.	NRT	A comparison of achievement of local black to the achievement of black students throughout the nation.	Central office
To compare achievement of one local school's student subgroup (e.g. sex, race, or age) to achievement of another such subgroup in a given year to determine the equity of educational outcomes.	NRT	A comparison of achievement of black and white students in local schools to determine and monitor any gap in achievement.	Central office, principals
To assess the extent to which students in a single grade level (at district, building, or classroom level) have mastered the essential objectives of the school system's curriculum.	CRT	A comparison of difference between results of September and May criterion-referenced tests to determine the extent to which 3rd graders at a given school attained 3rd grade objectives in reading.	Teachers, principals, central office
To assess the extent to which a given student is learning the essential objectives of the school system's curriculum and, subsequently, to adjust instruction for that student.	CRT	The use of the results from the September and January criterion-referenced tests as one indicator to help determine if a student is properly placed in an instructional group.	Teachers, principals, parents

[1] This chart was prepared by Prince George's County (MD) Public Schools.

Table 1. Appropriate Uses of Norm-referenced and Criterion-referenced Tests (continued)

Purpose	Test	Example	Primary Users
To track achievement of cohort of students through the system or area to determine the extent to which their achievement improves over time.	CRT	An examination of progress of all 3rd graders in system, administrative area, or school from one year to the next.	Central office, principals
To track achievement of cohort of students in a given school to determine the extent to which they learn essential objectives of school system's curriculum as they go from grade to grade.	CRT	The use of May criterion-referenced tests (or perhaps gains from September to May), to follow the progress of children over time in terms of the extent to which they learned the curriculum from one year to another.	Principals, teachers

Some Measurement Concepts[1]

WHAT TYPES OF TEST SCORES ARE THERE?

Different types of scores provide different types of information and serve different purposes. You should understand the different types of scores before you can use them or select scores that are most appropriate for your needs.

In this section, we define these types of test scores:

- *raw scores,*
- *total percentage correct scores,*
- *object mastery scores,*
- *percentile scores,*
- *stanine scores,*
- *grade equivalent scores,*
- *standard scores, and*
- *normal curve equivalent scores*

and explain the advantages and disadvantages of each. In the next section, we discuss how to use them.

Remember that test scores reflect only what was measured on a particular test (its domain). For example, scores on the Iowa Tests of Basic Skills (ITBS) test of mathematics achievement reflect only the combination of skills tested by the ITBS. Scores on other mathematics tests are comparable to the extent that their domains are comparable.

Raw scores

Raw scores indicate the number of items a student answers correctly on a test. For students who take the same test, it makes sense to compare their raw scores. If one third grade student answers 12 of 25 items correctly and another answers 16 correctly, then we likely will conclude the second student knows the content of that test better than the first.

Because the number of items varies between tests and because tests vary in difficulty, raw scores have little value in making comparisons from one subject to another. Suppose a third grade student answers 12 out of 25 items correctly on a mathematics test and 16 out of 25 items on a reading test. Some people may assume that the student is better in reading than in mathematics. However, we really know nothing about relative performance in the two different areas because the mathematics test may be much harder than the reading test.

How Are Raw Scores Distributed?

As an example of how raw scores are usually distributed over the population, let's look at a national sample of 2,000 students.

---

[1] Written by Lawrence Rudner

If you give a 25-item mathematics test to a large number of students, you will typically find the largest number of students have scores around the average, or mean, and the number of students with a given raw score decreases the further you get from the mean.

Figure 1 illustrates a hypothetical number of students with each test score.

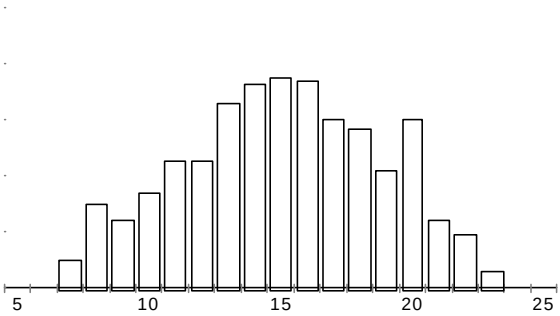


Figure 1. Raw Scores

The distribution of test scores shown in Figure 1 can be modeled mathematically using the familiar bell-shaped "normal" curve.

In the normal curve shown in Figure 2, the *y* axis shows the relative proportion of students and the *x* axis shows total raw score. The curve is used to approximate the proportion of students who would have a given total score.

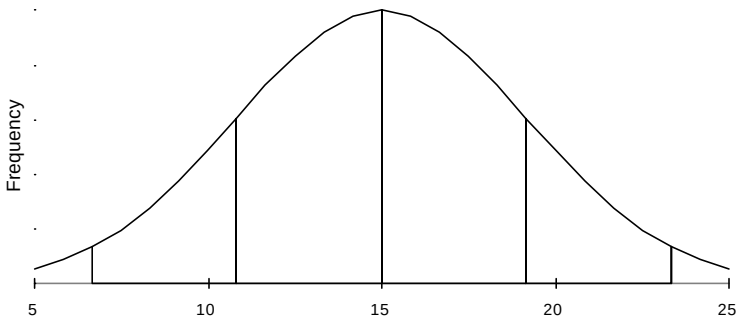


Figure 2. The Normal Curve

The normal curve is only a mathematical model that shows a relationship between two variables -- **test scores** and **proportion of students**. Actual scores never perfectly match the model. Nevertheless, the model is close to reality and gives good practical results. The same relationship between test scores and proportion of students holds for a wide number of tests.

Test developers use the model of the normal curve in developing and norming tests. In this guide, we use it to show similarities between different types of normative test scores -- test scores that describe individual student performance in comparison to the actual performance of a large group of students.

Two statistics are helpful in discussing test score distributions:

- the *mean* and
- the *standard deviation*.

The *mean* is frequently called the *average* score. You compute the mean by adding all the scores then dividing the sum by the total number of scores.

A *deviation* score is *how far away the score is from the mean*. For example, on a test with a mean of 15, a score of 20 deviates 5 points from the mean. The deviation score alone does not tell you whether this is a big difference or not. Rather, the *standard deviation* gives you a framework for interpreting this test score variability. You compute the standard deviation by taking the square root of the average of all the squared deviations. You can interpret standard deviation as an average distance that the scores deviate from the mean.

What are the advantages of raw scores?

- They are easy to compute.
- One of the most accurate ways to analyze a student's gains in achievement is to compare the raw scores from two administrations of the same test.

What is the limitation of raw scores?

Raw scores do not contain a frame of reference for indicating how well a student is performing.

Total percent correct scores

Total percent correct scores tell you the percentage of items that a student answers correctly out of the total number of items on a test. Like raw scores, total percent correct scores do not reflect varying degrees of item and test difficulty. They are of limited value in making comparisons.

Note that total percent correct scores are NOT the same as percentile scores. (We discuss percentile scores later in this section.)

What are the advantages of total percent correct scores?

- They are easy to compute.
- They adjust for differing numbers of items.

What are the limitations of total percent correct scores?

- They do not adjust for differing test difficulties.
- They do not contain a frame of reference for indicating how well a student is performing.
- They can mislead teachers, students and others into thinking the percent correct a student receives is the percent of the content the student knows or can do.

Objective percent correct scores

Objective percent correct scores tell you the percent of the items measuring a single objective that a student answers correctly. Because objectives and items can vary in difficulty, this score is of limited value for determining whether a student has mastered a learning objective. Indeed, the objective percent correct scores is really a percent correct score for a reduced test domain (reduced to a single objective).

You might interpret the objective percent correct score in relation to an **expected** objective percent correct. Expectations are sometimes based on curricular goals, last year's performance, or national averages. But, since different collections of test items will not be equivalent in difficulty, comparing a student's objective percent correct with another student's or with an expectation should only be done when the items are identical or equivalent.

Expectations can be used to convert objective percent correct scores to *objective mastery scores*. When the expectation is met or exceeded, the **objective is mastered**. Conversely, when the score is lower than expected, the objective is not mastered.

For example, suppose a test contains eight whole-number addition problems and a student answers seven of them correctly. That student's objective percent correct score is 87.5%. If you feel that answering, say,, three out of every four questions correctly reflects mastery, then this test score indicates that the student has mastered the objective.

What are the advantages of objective mastery scores?

- They are easy to compute.
- They adjust for differing numbers of items per objective.
- They help you diagnose specific individual strengths and weaknesses.
- They provide a skill-based approach to classroom grouping and school-based curricular emphasis.

What are the limitations of objective mastery scores?

- They require a fairly large number of items (usually more than ten) for each objective. The fewer items there are per objective, the greater is the likelihood of mistaking masters from non-masters and vice versa.
- Expectations are not always easy to define. The national average is not always a good basis for determining expectation.
- They do not indicate the degree or level of skill that the student has attained; they only indicate the status of mastery or non-mastery.

Percentile scores (ranks)

Percentile scores tell you the percent of students in the norming sample whose scores were at or lower than a given score. Percentile scores are among the most commonly reported scores and are best used to describe a student's standing in relation to the norming group at the time of testing. For example, if a student's score is in the 80th percentile, then that student scored equal to or higher than 80% of the students who took the test when the test was normed.

Note that although percentile scores are reported in increments of one hundredths, they are not completely accurate. When you use percentiles, you should pay attention to the *confidence bands* that the test publisher provides.

Confidence bands represent the **range of scores** in which a student's true score is likely to fall. For example, although a student's score on a particular test may be at the 86th percentile, it is likely that if the student took the same test on a different day, the new score would vary slightly. Accounting for random variations, that student's true achievement may fall somewhere within a range of scores, for example, between the 81st and 89th percentiles.

Percentile units are used to report an individual student's score; they should not be averaged to describe groups. Percentile units cannot be subtracted to compute gains because differences in percentile scores are not constant across the entire scale. For example, getting an additional two items correct can greatly increase a percentile rank for an average student. Yet the score increase from the same two items may not result in any percentile change for students of very above average achievement. Score gains increase percentile ranks more in the middle of the range than toward the extremes. (See Figure 3.)

How are percentile scores distributed?

Figure 3 shows how percentile scores are distributed when raw scores are distributed normally. The *y* axis shows the proportion of students and the *x* axis shows the percentile score. Vertical lines have been drawn to indicate each standard deviation unit. Note that the percentile scores are not evenly distributed on the *x*-axis. If they were evenly distributed, then the proportions graphed on the *y*-axis would all be the same; each proportion would be 1%!.

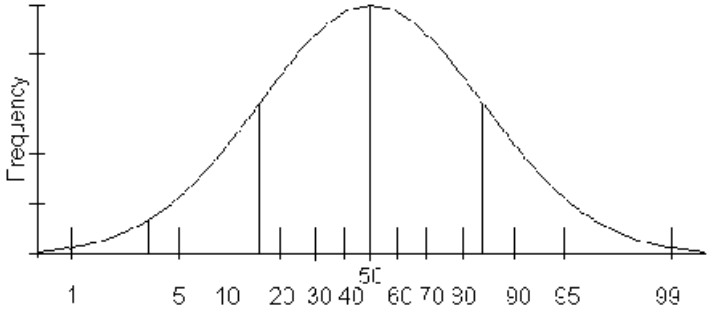


Figure 3. Percentile Score Distribution

Notice that percentiles are more "spread out" at the ends of the figure. For example, the raw score difference between the 95th and 90th percentile is greater than the difference between the 55 and 50th percentile. This happens because a student needs to answer more items correctly to move from the 90th to the 95th percentile than is necessary to move from the 50th to 55th percentile. Therefore, scores are clustered around the mean. It is because of this difference that you should not add, subtract, or average percentiles.

What are the advantages of percentile scores?

- They show how students rank in relation to the national or local average.
- They are easy to explain.

What are the limitations of percentile scores?

- They can be confused with total percent correct scores.
- They are not as accurate as they appear to be.
- They are often used inappropriately to compute group statistics or to determine gains.
- They are frequently misunderstood.

Stanine scores

Stanine is short for *standard nine*. Stanine scores range from a low of 1 to a high of 9 with:

- 1, 2, or 3 representing **below average**
- 4, 5, or 6 representing **average**
- 7, 8, or 9 representing **above average.**

If a student achieves a stanine score that is below average in a particular area, the test has revealed an area in which the student may need to improve -- or at least it reveals an area in which the student is weak when compared to other students who took the test. If the student achieves an average stanine score, the test has revealed that the student performed

at the same level as most of the other students who took the test. Similarly, if the student achieves a stanine score that is above average, the test revealed that the student performed better in that area than most of the other students who took the test.

Stanines are frequently used as a basis for grouping students. For example, an advanced mathematics class may enroll students in the 9th, 8th, and sometimes 7th stanine.

How are stanine scores distributed?

Figure 4 shows how stanine scores are distributed when raw scores are distributed normally. The *y* axis shows the proportion of students and the *x* axis shows the stanine score. Vertical lines have been drawn to indicate each standard deviation unit. Stanine 5 represents ½ a standard deviation (sd) around the mean. Stanines 2, 3, 4 and 6, 7, and 8 also represent the same raw score difference (½ sd). Stanines 1 and 9 represent all the scores below -1.75 sd and above +1.75 sd, respectively.

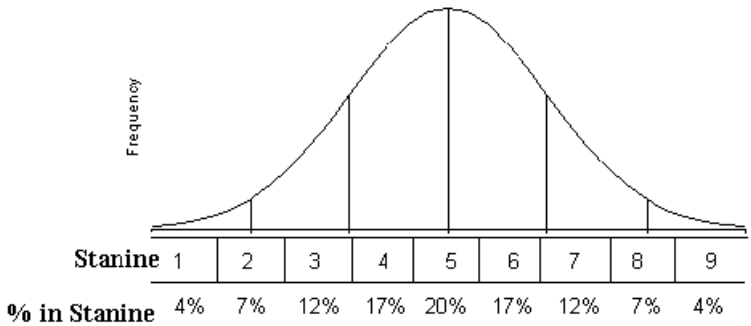


Figure 4. Stanines

Stanine scores are normalized. This means they will be distributed normally whether or not the original test was normally distributed. This is accomplished by assigning the highest and lowest 4% of the test scores to stanine 9 and 1, respectively; the next highest and lowest 7% to stanines 8 and 2; the nexthighest and lowest 12% to stanines 7 and 3, the next highest and lowest 17% to stanines 6 and 4, and the middle 20% to stanine 5. The percentages were chosen to approximate a normal distribution.

District test results can be reported by showing the percent of district students who fall in each stanine computed based on a national norming group.

What are the advantages of stanine scores?

- They show the standing of students in relation to the national or local average.
- They are relatively easy to explain.
- They can be used to group students into ability groups.

What are the limitations of stanine scores?

- They should not be used in computing group statistics or in determining gains.
- They give only very general indications of a student's relative standing in a particular content area.

Grade equivalent scores

Grade equivalent scores use a scale based on grade levels and months to estimate how well students perform. These scores reflect the median score of students across several grade levels during the month the test was normed. For instance, the median test score for first graders in the seventh month of the school year (April) would convert to a score of 1.7, for second graders the score would be 2.7, for third graders the score would be 3.7, and so forth.

Grade equivalent scores are often misunderstood. For example, if a fourth grader received a grade equivalent score of 7.0 on a fourth grade reading achievement test, some people may assume that the fourth grader has mastered seventh grade material. However, the score actually means that the fourth grader reads fourth grade material as well as the typical beginning seventh grader (in September) would read the same fourth grade material.

As with percentile scores, you should use grade equivalent scores only to describe a student's standing in relation to the norming group at the time of testing. You should not average grade equivalent scores to describe groups, and you should not subtract them to compute gains.

As with differences in percentile scores, differences in grade equivalent scores do not mean the same thing across the entire scale.

How are grade equivalent scores distributed?

Figure 5 shows an example of how grade equivalent scores are distributed when raw scores are distributed normally. The *y* axis shows the proportion of students and the *x* axis shows the grade equivalents. Vertical lines have been drawn to indicate each standard deviation unit.

Note that this is just an example, because grade equivalent scores are not defined by the model but rather by the actual performance on the test by students in higher and lower grade levels.

Notice that relatively few correct responses translate to large differences in grade equivalent scores for students who achieve very high and very low scores. Because of this, grade equivalent scores do not estimate group ability well and you should not use them to evaluate gains over time.

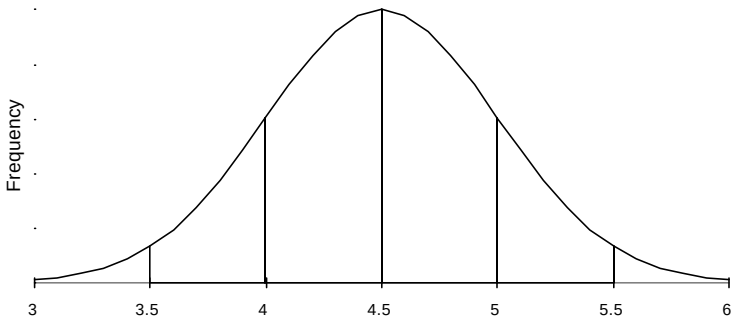


Figure 5. Grade Equivalent Scores

What is the advantage of grade equivalent scores?

Grade equivalent scores are expressed in grade-level values that are familiar to parents and teachers.

What are the limitations of grade equivalent scores?

- They are frequently misunderstood and misinterpreted.
- They have low accuracy for students who have very high or very low scores.
- They should not be used for computing group statistics or in determining gains.

Standard scores

Standard scores tell you how much students' scores deviate from a mean. Almost all of the companies that publish achievement tests will give you standard scores. However, they often use different names -- such as *growth scale values, developmental standard scores,* and *scaled scores* -- and different units to report the scores. Thus, a scaled score of 110 on one test may not be the same as a scaled scores of 110 on another.

The main advantage of standard scores is that they give you an equal interval unit of measurement. As a result, you can use them to compute summary statistics, such as averages and gains, if all the students you compare took the same test. A two-point difference between standard scores means the same difference, no matter where a students falls within the range of scores (unlike percentile and grade equivalent scores).

As we noted, the scales used for standard scores differ among test publishers and among content areas. As a result, you cannot usually use these scores to compare results on different tests.

How are standard scores distributed?

Figure 6 shows how standard scores are distributed on the a hypothetical test when raw scores are distributed normally. Here the raw scores have been translated to a scale with a mean of 100 and a standard deviation of 10. The *y* axis shows the proportion of students and the *x* axis shows the standard score. Vertical lines have been drawn to indicate each standard deviation unit.

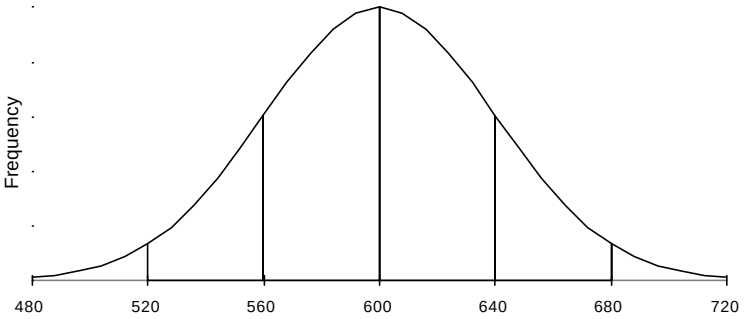


Figure 6. Standard Scores

Note that the intervals in Figure 6 are equal in size. This feature makes standard scores and scores based on standard scores the statistic of choice when reporting group averages and changes over time.

What are the advantages of standard scores?

- They are linearly related to raw scores and thus have many of the advantages of raw scores.
- They show relative performance of a student within a group.

What are the limitations of standard scores?

- They can be confusing to parents and teachers unless they are converted to percentile scores.
- They have no intrinsic meaning, unless the scale is commonly understood because it is used frequently. For example, the Scholastic Assessment Test for college admissions uses a standard score with a mean of 500 and a standard deviation of 100.

Normal curve equivalent scores

Normal curve equivalent scores were originally developed to analyze and report gains in compensatory programs for educationally disadvantaged students. These scores have a mean of 50 and a standard deviation of approximately 21. This results in a scale with 99 equal interval units.

A normal curve equivalent score of 50 represents the national average of any grade level at the time of year the test was normed. A score of 30 is always the same distance below grade level, regardless of the level tested, and is twice as far below grade level as a score of 40.

Normal curve equivalent scores are similar in their range to percentile scores, but they have statistical properties that allow them to be used to compute summary statistics and gain scores.

How are normal curve equivalent scores distributed?

Normal curve equivalents are normalized scores (see the discussion of stanines above). Figure 7 shows how normal curve equivalent scores are distributed. The *y*-axis shows the proportion of students and the *x*-axis shows the score. Vertical lines have been drawn to indicate each standard deviation unit.

Because normal curve equivalents are a type of standard score, they have the same statistical properties as standard scores. Normal curve equivalent intervals are of equal size and these scores can be used to compute group statistics.

What are the advantages of normal curve equivalent scores?

- They allow you to compare the performance of students who take different levels or forms of the same test within a test battery.
- They allow you to draw comparisons across subject matter for the same student.
- They can be used to compute meaningful summary statistics.

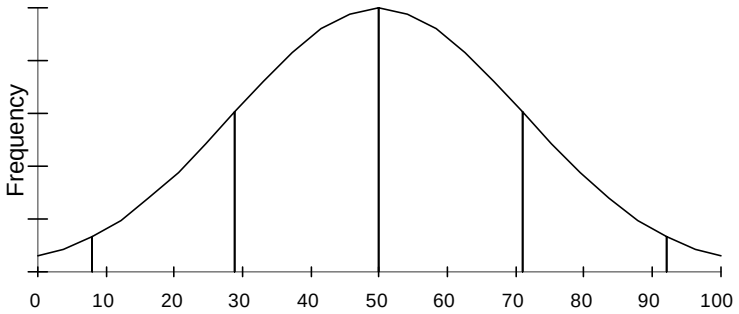


Figure 7. Normal Curve Equivalent Score

- They can be used to evaluate gains over time.
- They can be used to combine data from different tests.

What is the limitation of normal curve equivalent scores?

Normal curve equivalent scores do not give you easily understood information about an individual student's achievement level, unless they are compared to another value or are converted to a percentile score.

HOW SHOULD YOU USE TEST SCORES?

Interpreting norm-referenced test scores

Normative test scores -- stanines, percentile scores, scaled scores, and grade equivalent scores -- measure **an individual** student's achievement in relation to the achievement of **one or more large groups** of students who took the same test. The comparison group may be composed of other students in your district or of students from a nationally representative sample. Thus, scores on norm-referenced tests are meaningful only in relationship to a comparison group.

Your school or district is not completely like the normative group. No district is. In many cases, the differences are minor and inconsequential. However, in other cases, schools can be so different that the national norms provided by the publisher do not accurately reflect school performance. Norms become less meaningful as your students and your testing program become more unlike the standardization sample.

If your students are tested at a different time of the year than the norm group was tested, the interpretation of the percentile score is unclear. For example, the CAT is normed in October. That means that you must give it in October to make your students' scores most meaningful. If you give the CAT in January, you cannot know if a student who scores in the 55th percentile is above or below average when compared to grade-level peers. (See the Appendix called *Communicating a complete report card for your school* for a list of the many ways in which your students, schools, and district may be different from the normative sample.)

Many of these differences can seriously affect your scores. This does not mean the national norms are useless; it means that you must evaluate the norms in perspective. Some publishers extrapolate norms so they are based on the week the test was given, for example. Norms give you an index of how well students perform on certain tasks -- tasks the test publishers have identified as representing the skills taught to the comparison group at the time the test was developed.

Norm groups are used at varying points in time but their data are actually historical. Scores that are above average, for example, may be only above the average of students in the norm group who were tested four years ago. They may not be above today's average for a similarly defined group..

The comparative baseline of norm-referenced tests is a powerful tool. In addition to worrying whether your Chapter 1 students are learning basic skills, for example, you probably are also interested in how well they are doing in relation to the nation. Although

your students may not be like the nation at large, they are going to be competing for jobs and educational opportunities against a wide range of other students.

While national averages give you a baseline, you must establish your own expectations and goals considering your particular community and curriculum. For example, it would be somewhat misleading for you to report above average scores for a magnet school that selects students based on academic achievement. In this case, you would be better off reporting on the gains or specific achievements of the students who are in the program.

References and Recommended Reading

Anastasi, Anne (1988). *Psychological Testing*. New York, New York: MacMillan Publishing Company.

Lyman, Howard B. (1993). *Test Scores and What They Mean*. Boston: Allyn and Bacon.

McMillan, James H. (2001). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Publishing Company.

.

Using State Standards and Assessments to Improve Instruction[1]

Today many states around the country have curriculum standards, and state developed assessments to monitor the implementation of those standards. Most state standards define expected outcomes, that is, what students need to know and be able to do, but do not mandate specific strategies or pedagogy used by local districts. Elementary, middle and high school students around the country take at least one state mandated test during their school career. However, 35 out of 50 states do not require teachers take a course, or demonstrate competency, in the area of assessment. Hence, teachers generally have limits to their knowledge of how to design and use tests and assessment tools. Richard Stiggins (1999) wrote, " It is time to rethink the relationship between assessment and effective schooling."

It is possible for teachers and administrators to use state content and process standards, test specifications, curriculum frameworks, sample questions, educational research, and exemplar papers to improve instruction and classroom tests and assessment procedures, but limited understanding puts constraints on this use. Researchers Paul Black and Dylan Wiliam (1998) stated standards are raised only by changing what happens in the classroom, beginning with teachers and students. These researchers go on to say that a large body of evidence suggests that attention to formative assessment is a vital feature of classroom work and the development of it can raise standards.

This article describes a program used by two educators to help teachers improve instruction through a deeper understanding of state standards and test specifications. Any teacher or administrator in any state can use the process outlined in this article. Specific examples were developed using the New Jersey Core Curriculum Content Standards and that state's fourth grade mathematics test.

DEVELOPING A KNOWLEDGE BASE

Understanding how standards-based state tests are constructed is the first step in being able to use them to guide and improve instruction. A test is essentially a sample of questions or activities that reflect a large body of knowledge and mental processes associated with an academic subject area. It is highly impractical to design a test that includes all of the problems that a student could ever do in each content area. Therefore, state test are samples of possible questions from each area. All state tests are limited samples of what students are required to know in areas such as language arts, mathematics, science, etc. There are large numbers of questions that can appear on future forms of these instruments. A teacher would not be able to address all the possible questions, nor should the teacher attempt that task. However, school districts and teachers should endeavor to understand the delineation of each subject area.

School districts are under pressure to perform well on state tests and often use a test preparation strategy of giving students sample tests from commercially prepared

---

[1] Written by Christopher Tienken and Michael Wilson.

workbooks or state released items to get ready for state tests. Although this is one strategy that can be useful for providing general information regarding student strengths and weaknesses as related to the samples, it should not be the only method used by teachers. The strategy itself, does little to educate teachers about how to use and understand state tests, standards, and test specifications. This article recommends a three-part process for developing an understanding of state assessments and using that understanding to improve instruction. That process is delineation, alignment, and calibration.

DELINEATION

Delineation is the first component needed to understand any standards based test. It is the process of thoroughly identifying all aspects of a particular subject domain; the aspects are also known as dimensions. Delineation involves the use of state testing documents that describe each content area of the assessment. The documents usually include test specifications, specific skill cluster information, subject area frameworks, assessment examples and exemplars, and the state standards. Delineation requires an examination of these documents for assessment dimensions such as content, cognitive level and complexity. A thorough delineation might also include analysis of the test format, motivation, the difficulty level of the questions, and related affective characteristics of the subject area.

Thoroughly examining state standards and test specifications is a way to begin delineation. The New Jersey Standards include macro or big picture statements and cumulative progress indicators that provide details about general performance expectation. The State's test specifications are particularly helpful because they go further and break the Standards down into two distinct types. Knowledge specifications describe the specific processes and content that all students must know by the end of fourth grade. Some would call these content standards. Problem solving specifications describe what students should be able to do with the content knowledge. They are also known as process standards. The following example is excerpted from the 4th grade New Jersey mathematics standards and test specification manuals.

Macro Standard 4.1: All students will develop the ability to pose and solve mathematical problems in mathematics, other disciplines, and everyday experiences.

Cumulative Progress Indicator 4.1.2: Recognize, formulate, and solve problems arising from mathematical situations and everyday experiences.

> Test Specification Manual - Cluster IV Discreet Mathematics:
>
> Knowledge (content standards): Students should have a conceptual understanding of: Tree diagram
>
> Problem Solving (process standards): In problem solving settings, students should be able to: Draw and interpret networks and tree diagrams

After reviewing the 4th Grade New Jersey Core Curriculum Content Standards and test specifications for mathematics, a teacher would be able to identify seven distinct mathematics strands or dimensions. Those strands are Numeration and Number Theory,

Whole Number Operations, Fractions and Decimals, Measurement/ Time/ Money, Geometry, Probability/Statistics, and Pre-algebra. Figure 1 represents the content delineation of the domain of mathematics after a team of 4th grade teachers examined the New Jersey Core Curriculum Content Standards, 4th grade state test specifications, and the local curriculum.

Mathematics Domain

Numeration/Number Theory	Whole Number Operations
Fractions/Decimals	Measurement/Time/Money
Geometry	Pre-algebra
Probability/Statistics	

(Delineated Strands / Dimensions)

(Figure 1 –A delineation of the domain of Mathematics)

Working through the different dimensions associated with the delineation process helps to increase teacher and administrator understanding of each content area and its relationship to the standards, classroom instruction and assessment.

The following activities can begin once teachers and administrators specify all of the subject area dimensions:

- selecting and designing classroom assessments and practice questions
- revising and designing curriculum that is congruent with the content identified in the state standards and the district's delineation of the state designed exams
- designing teacher training using instructional techniques that support these dimensions

A closer look at the 4th grade New Jersey Core Curriculum Content Standards and test specifications for mathematics reveals an emphasis on performance and the use of mathematics to solve open ended and word problems. The test specifications for that exam imply that the mathematics test questions are primarily composed of problem solving tasks. Therefore, it is safe to assume that test questions will require thinking in the application, analysis, and perhaps synthesis and evaluation levels of cognition.

ALIGNMENT

During the alignment phase, administrators and teachers work to identify, analyze, generalize, and describe the links between the various elements associated with the subject area previously delineated and the sample questions selected for practice or classroom activities to assess student progress. The sample questions and student assessments can be derived from several sources including state released test items, commercially manufactured

test preparation materials, or teacher made activities. Teachers and administrators examine linkages in the materials, organization, textbooks, instructional strategies and other elements described in the curricula and used in daily instructional activities to ensure consistency with the district's delineation of the state assessment.

Using and understanding the test specifications become even more important at this stage. Let's imagine that a pair of 4th grade teachers recently completed a delineation of the mathematics domain and identified their next unit of study. The unit centered on Standard 4.1.1 and the test specification listed below. Reviewing the prior example from the test specification manual and Cluster IV the teacher would complete several alignment tasks:

Test Specification Manual - Cluster IV Discreet Mathematics:

Knowledge (content standards): Students should have a conceptual understanding of: Tree diagram

Problem Solving (process standards): In problem solving settings, students should be able to: Draw and interpret networks and tree diagrams

Tasks:

1 Review classroom resources, curriculum, textbooks, teacher activities, student thinking strategies and tests to ensure that the above test specifications and macro standards are addressed on the knowledge and problem solving level. Do the teacher resource materials and classroom instruction address the proper skills?

2 Review the above factors to ensure congruency between the level of difficulty required by the standards and specifications, and the difficulty of the actual teacher resources and activities. Do the teacher's tests, lessons, activities etc., match the difficulty level required by the standards and specifications?

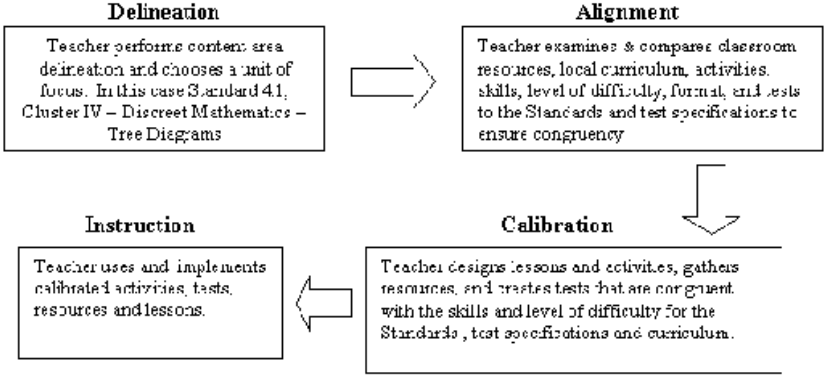
3. The teacher must also consider format. Although less important than skills and difficulty, the teacher resources, activities, and tests should familiarize the students with state test question formats.

4. Teachers must align classroom assignments and activities to the subject area delineation to ensure congruency.

CALIBRATION

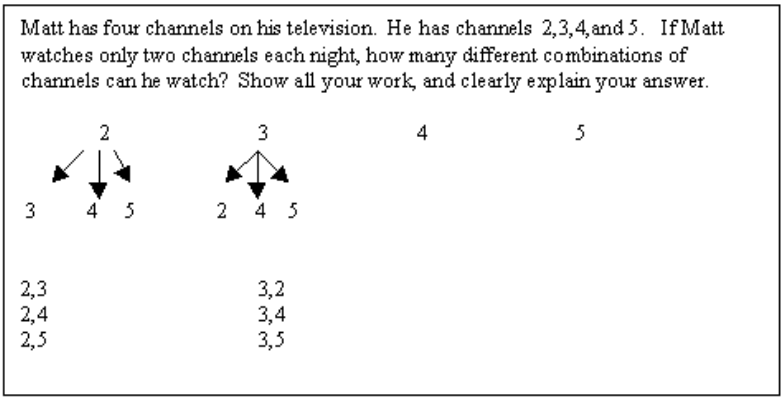
After completing the delineation and beginning the alignment processes, calibration begins. Calibration is the act of conducting communications and interactions with teaching staff based on the information identified in delineation and used in alignment. The calibration process ensures that the conceptualization of content, cognitive process, complexity, formats, etc. is consistently understood for each subject area. Calibration, in its simplest form, is designing classroom instruction, activities and assessments that are

congruent with content area delineation and alignment. Using the prior mathematics vignette as an example, one can begin to see how the process takes place. Figure 2 represents the sequence of events leading up to calibration.



(Figure 2. Delineation, Alignment, and Calibration Flow of Events)

Imagine that a 4th grade teacher completed delineation and alignment and discovered that her/his program was missing a unit on discreet mathematics. That teacher would develop objectives related to understanding, using, and interpreting tree diagrams. Figure 3 is a sample activity / test question created by 4th grade teacher Terry Maher to begin addressing the aspect of discreet math noted in the Cluster IV test specification.



(Figure 3: A sample activity/ test question)

Calibration is any action that helps teachers design activities and construct assessments based on the dimensions of state assessments and standards. This process helps to foster a collective understanding and agreement of the dimensions and domains of each content area. It should be a team effort based on group inquiry.

USING SCORE REPORTS TO IMPROVE CALIBRATION

As teachers gain a better understanding of how student work reflects the standards and test specifications through delineation, alignment and calibration, their efficiency and accuracy at identifying which students are meeting the standards should increase. Herein lies the usefulness of score reports. State test score reports sort students into categories of varying proficiency. For example, a student who scores partially proficient, proficient, or advanced proficient on a state language arts test may also show some congruency in the level of achievement in his/her well-aligned school work and classroom assessments. As teachers become better calibrated, they will be able to answer questions such as: Is the student showing partial proficiency, proficiency, or advanced proficiency on class assessments? If not, why? Is the difficulty level of the class work comparable to the state exam? What can I do to help this student meet the state standards? Is my program meeting the standards?

PREDICTING OUTCOMES

Teachers can reflect upon their level of calibration accuracy by attempting to predict student results on state assessments. This type of exercise acts as an extension to the calibration process and can provide teachers with a way to get a very general sense of their level of calibration. Teachers should be aware that there would not be 100% agreement between a student's performance on well-calibrated classroom tests and state assessments based on many factors of test design. This process is meant to compliment the calibration exercises and provide the teacher with extra data regarding their calibration exercises.

To begin the prediction process, the teacher uses a list of the students taking the test. Beside each name, the teacher enters a predicted score level. When the state assessment scores arrive, the teacher can compute the level of accuracy as shown below.

Name	Prediction	Score
Allan	Proficient	Adv. Proficient
Ann	Proficient	Proficient
Tamika	Adv. Proficient	Proficient
Bronson	Partial Proficient	Partial Proficient

The list above shows a 50% level of success in the predictions made. The teacher making the predictions returns to each student's work and compares the successful predictions with the unsuccessful ones to gain a better idea of how the assessment performances reflect the aligned student work. Student work associated with actual test scores can form the basis for subsequent calibration discussions. Student work connected to state assessment score levels can also function as scoring examples that students refer to when judging their own work.

FINAL THOUGHTS

The process outlined in this paper is very different from the idea of using testing materials and example tests to teach specific items on state assessments. Although there is a place for such strategies, this article suggests that it is more important for the teacher to understand the entirety of each subject area, and where state test content fits within each of these areas. Teachers must teach toward an understanding of the subject areas while they

align and calibrate their classroom activities, resources, tests, and instruction with the specifications and skills required by each state's standards. There is a distinct difference between traditional notions of test preparation and aligning and calibrating instruction and assessments with the content, cognition, difficulty, and format of state assessment instruments, specifications, and standards. The aim is to ensure that teachers understand, and calibrate their classrooms with respect to the entire process and do not simply focus on how to answer specific types of test questions.

The questions will change, but the underlying skills and concepts will not. One must be careful not to wallow in the mire of test prep. As educators, we are trying to link the classroom activities to the standards and skills set by the state. Delineation, alignment, and calibration are academic endeavors that demand unending commitment. Do not expect to accomplish alignment or calibration at an in-service day, or even during the course of a school year. This ongoing process requires constant attention. The administration must provide the time and resources to conduct frequent calibration meetings to examine such things as classroom work and student assessment samples. Beware, it is easy to fall out of alignment and calibration and into test prep.

References

Black, Paul, and Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment, *Phi Delta Kappan* October, pp. 139-148.

Stiggins, R. (1999). Assessment, student confidence, and school success, *Phi Delta Kappan* November, pp. 191-198.

Related Works and Readings

Neill, D. (1997 September). Transforming student assessment, *Phi Delta Kappan,* pp.35-36.

Sadler, D. (1989). Formative assessment and the design of instructional systems, *Instructional Science*, vol. 18, 1989, pp. 119-44.

Schafer, W. & Lissitz, R. (1987). Measurement training for school personnel: Recommendations and reality, *Journal of Teacher Education*, vol. 38. 3, pp. 57-63.

Stiggins, R. & Conklin, N. (1992). In teachers' hands: Investigating the practice of classroom assessment. Albany: State University of New York Press.

Preparing Students To Take Standardized Achievement Tests [1]

The public often favors accountability in education and believes that holding teachers responsible for students' achievement will result in better education. Many people assume that the best data about students' levels of achievement come from standardized achievement tests. Although scores from these tests are undoubtedly useful for accountability purposes, educators recognize that such data have some limitations.

TEACHING TO THE TEST

One major concern about standardized achievement tests is that when test scores are used to make important decisions, teachers may teach to the test too directly. Although teaching to the test is not a new concern, today's greater emphasis on teacher accountability can make this practice more likely to occur.

Depending on how it is done, teaching to the test can be either productive or counterproductive. Therefore, you need to carefully consider how you prepare students to take standardized achievement tests.

At some point, legitimate teaching to the test can cross an ill-defined line and become inappropriate teaching of the test (Shepard and Kreitzer, 1987). Educators may disagree about what specific activities are inappropriate. However, it may be useful to describe a continuum and to identify several points located along it.

SEVEN POINTS ON THE CONTINUUM

Mehrens and Kaminski (1989) suggest the following descriptive points:

1. giving general instruction on district objectives without referring to the objectives that the standardized tests measure;
2. teaching test-taking skills;
3. providing instruction on objectives where objectives may have been determined by looking at the objectives that a variety of standardized tests measure (The objectives taught may or may not contain objectives on teaching test-taking skills.);
4. providing instruction based on objectives (skills and subskills) that specifically match those on the standardized test to be administered;
5. providing instruction on specifically matched objectives (skills and subskills) where the practice or instruction follows the same format as the test questions;
6. providing practice or instruction on a published parallel form of the same test; and
7. providing practice or instruction on the test itself.

Mehrens and Kaminski suggest that:

- Point 1 is always ethical and Points 6 and 7 are never ethical.
- Point 2 is typically considered ethical.

[1] Written by William A. Mehrens

Thus, the point at which you cross over from a legitimate to an illegitimate practice on the continuum is somewhere between Points 3 and 5. The location of the point changes depending on the inferences you want to make from the test scores.

WHAT YOU CAN INFER FROM TEST SCORES

"The only reasonable, direct inference you can make from a test score is the degree to which a student knows the content that the test samples. Any inference about why the student knows that content to that degree...is clearly a weaker inference..." (Mehrens, 1984, p. 10).

Teaching to the test alters what you can interpret from test scores because it involves teaching specific content. Therefore, it also weakens the direct inference that can be reasonably drawn about students' knowledge. Rarely would you want to limit your inference about knowledge to the specific questions asked in a specific format. Generally, you want to make inferences about a broader domain of skills.

Further complicating matters, many people wish to use test scores to draw indirect inferences about why students score the way they do. Indirect inferences can lead to weaker and possibly incorrect interpretations about school programs.

Indirect inferences cannot possibly be accurate unless the direct inference of student achievement is made to the correct domain. Rarely does one wish to limit the inference about knowledge to the specific questions in a test or even the specific objectives tested. For example, if parents want to infer how well their children will do in another school next year, they need to make inferences about the broader domain and not about the specific objectives that are tested on a particular standardized test. For that inference to be accurate, the instruction must not be limited to the narrow set of objectives of a given test. Thus, for the most typical inferences, the line demarking legitimate and illegitimate teaching of the test must be drawn between Points 3 and 4.

While in my view it is inappropriate to prepare students by focusing on the sample of objectives that happen to be tested, you can undertake appropriate activities to prepare students to take standardized tests.

APPROPRIATE ACTIVITIES TO PREPARE STUDENTS

Ligon and Jones suggest that an appropriate activity for preparing students for standardized testing is:

"one which contributes to students' performing on the test near their true achievement levels, and one which contributes more to their scores than would an equal amount of regular classroom instruction" (1982, p. 1).

Matter suggests that:

"Ideally, test preparation activities should not be additional activities imposed upon teachers. Rather, they should be incorporated into the regular, ongoing instructional activities whenever possible." (1986, p. 10)

If you follow the suggestion by Ligon and Jones, you might spend some time teaching students general test-taking skills. These skills would help students answer questions correctly if they have mastered the objectives. Without some level of test-taking skills, even knowledgeable students could miss an item (or a set of items) because they did not understand the mechanics of taking a test.

SUMMARY

Although the temptation exists to teach too closely to the test, teachers should not be pressured to do so. In fact, you should try to ensure that they do not do so.

The inferences you typically wish to draw from test scores are general in nature and will be inaccurate if you limit instruction to the actual objectives sampled in the test or, worse yet, to the actual questions on the test. However, it is appropriate to spend some instructional time teaching test-taking skills. Such skills are relatively easy to teach and should take up very little instructional time.

REFERENCES

Ligon, G. D. and Jones, P. (April l, 1982). Preparing Students for Standardized Testing: One District's Perspective. Paper presented at the annual meeting of the American Educational Research Association, New York.

Matter, M. K. (1986). "Legitimate Ways to Prepare Students for Testing: Being Up Front to Protect Your Behind." In J. Hall and F. Wolmut (eds.). National Association of Test Directors 1986 Symposia. (pp. 10-11). Oklahoma City, OK: Oklahoma City Public Schools.

Mehrens, W. A. (1984). "National Tests and Local Curriculum: Match or Mismatch?" *Educational Measurement: Issues and Practice*, 3, (3), 9-15.

Mehrens, W. A. and Kaminski, J. (1989). "Methods for Improving Standardized Test Scores: Fruitful, Fruitless or Fraudulent?" *Educational Measurement: Issues and Practices*, 8 (1), 14-22.

Shepard, L. A. and Kreitzer, A. E. (1987). "The Texas Teacher Test." *Educational Researcher*, 16(6), pp. 22-31.

The Politics of National Testing[1]

Most teachers are comfortable with developing and using tests for classroom purposes, whether to see how much students have learned, to provide a basis for grades, or to gain an understanding of individual students' strengths and weaknesses. And as state departments of education move forward with their testing programs, teachers are becoming increasingly familiar with tests used as measures of accountability. A third layer of testing arises on the national level and includes the National Assessment of Educational Progress (NAEP) and the voluntary national tests that have been under discussion since 1997. This chapter opens with a discussion of the political rationale behind national testing and provides an overview of the voluntary national testing movement. It then turns to a brief examination of NAEP, "the nation's report card," in both its national sample format and its state administration, which may be a backdoor to a true national test. Finally, action steps and resources are provided to enable teachers to take part in the ongoing debate about national testing.

Does the United States need to have some kind of test that every student in every state takes to demonstrate mastery of some agreed-upon body of knowledge and skills? Other countries do, but few have a decentralized, diverse education system similar to ours. A national test would require reaching agreement on several issues including:

- What the test should cover
- What format it should take
- At what point(s) it should be administered
- Who should administer it
- Whether and how any types of students (e.g., those in special education, those with limited English proficiency) should be exempted or accommodated
- When it should be administered
- How it should be scored
- What the consequences should be for doing well or poorly on it
- How it should fit in with existing state, school district, and classroom standards and assessments
- Who should participate in (and pay for) its development

It's important to note here that commercial test publishers have long offered achievement tests (e.g., the Iowa Test of Basic Skills, the California Achievement Test, the Terra Nova) that are administered to schools across the country and normed on national samples but are not in themselves national tests because individual schools, districts, or states decide for themselves whether to use them and which to select. The SAT is probably the most common test administered in the country, but it is intended to measure college-bound students' aptitude for college work, not academic achievement across a wide range of subjects for all students. And it has the ACT as competition.

The question of a true national test is a complicated one, and like many policy matters, it has strong political overtones. Over the past two decades, many politicians have moved

[1] Written by Carol Boston

Edited by W. Schafer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.

from an initial position of strong support for a national test as an accountability tool to opposition on the grounds that a national test would usher in a national curriculum and lead to further federal involvement in what should be a state and local matter. These policy-makers want states to set their own standards without interference from Washington; they see a national test as a de facto attempt by the federal government to dictate what is important for their students to learn.

On the other hand, some politicians seem to be less troubled by an expanded federal role in testing, but more suspicious about whether national testing would lead to genuine school improvement and higher student achievement or just sort out and penalize low-performing schools and the students in them, who are disproportionately low income and minority. They argue that until there is truly equal opportunity to learn for all students (with equal access to technology, highly qualified teachers, good facilities, and other learning inputs), testing is an empty exercise. Some politicians also fear that poor test scores might fuel discontent with the public school system and lead to more support for controversial initiatives such as vouchers for private school students.

Those in favor of national tests, on whatever side of the political fence they sit, point to:

- the value of having a common basis for comparing individual, school, district, and state performance;
- the importance of specifying content and performance targets to encourage high aspirations and achievement; and
- the potential motivating effect of tests if results are linked to hiring and college admissions decisions (Davey, 1992).

Those against national tests point to:

- the fallacy that tests alone lead to positive changes in education;
- lack of consensus about desired educational outcomes in various subject areas and the pitfalls of attempting to establish a national curriculum;
- limitations and biases inherent in testing, particularly multiple-choice tests but also performance-based ones.
- short-sightedness in not attempting to address the real equity issues related to the education of minority and low-income students (Davey and Neill, 1992).

VOLUNTARY NATIONAL TESTS

President Clinton's 1997 proposal to implement voluntary national tests offers a case study in the politics of national testing. In his State of the Union address in 1997, Clinton vowed to make national tests one of the centerpieces of his administration. President Clinton appealed to every state to "adopt high national standards, and by 1999, [to] test every 4th grader in reading and every 8th grader in math to make sure these standards are met." He viewed the national tests as having important individual and schoolwide benefits and consequences, stating, "Good tests will show us who needs help, what changes in teaching to make, and which schools to improve. They can help us to end social promotion. For no child should move from grade school to junior high, or junior high to high school until he or she is ready." (State of the Union address, February 5, 1997).

Six states and 16 large school districts signed onto the plan for voluntary national testing, though some urban districts, including Houston and Los Angeles, later retreated from their commitment to the reading test when it was determined that the test would be given in English only. The tests were to be put on a fast track, with implementation scheduled for spring of 1999.

While there were many unanswered questions about the national tests, the administration selected the subject areas and grades to be tested for sound reasons. It is widely held that students should be fluent readers by the end of third grade when the academic emphasis shifts from "learning to read" to "reading to learn." Mathematics proficiency is important to enable students to take higher math courses that are in turn considered "gatekeeping" courses for college entrance. The core question, however, and one which ultimately derailed the momentum of national testing, was whether the testing of individual students was a responsibility of the federal government or a preemption of states' rights.

Because the U.S. Department of Education is poorly regarded by many politicians, the first tactical move was to wrest control of the test from the Department of Education to the nonpartisan National Assessment Governing Board (NAGB), which provides policy guidance for the National Assessment of Educational Progress. Clinton concurred, and the Senate voted 87 to 13 in September 1997 to give NAGB control of both the policy and operations of the test. A few days later, however, the House voted 295-125 to ban any funds for the tests and made no mention of NAGB at all. A look at the rhetoric of the time is instructive. House Speaker Newt Gingrich spoke out strongly against imposing "Washington standards" on local schools and proposed instead that tax breaks and federal vouchers be put in place for parents who wished to pull their children out of public schools and put them in private schools (Houston Chronicle Interactive, September 9, 1997). Senate Majority Leader Trent Lott likened the national tests to stationing the IRS in the schools (Los Angeles Times, November 2, 1997).

Clinton signed a compromise bill in November 1997 that gave NAGB authority over contract work to develop the national tests, but stipulated that no FY 1998 funds could be used for pilot testing or field testing the instruments and instructed the National Academy of Sciences to evaluate the test and related testing issues. By February 1998, opposition to the testing led the House to vote 242-174 to approve a bill that would require Congress to "specifically and explicitly" authorize any test development in future fiscal years. Twenty-five Democrats joined 217 Republicans in the vote; only two Republicans voted against the measure. Marshall S. Smith, then acting deputy secretary of the U.S. Department of Education and point man for the voluntary national tests, viewed this vote as decisive evidence that the tests were in trouble. He told an audience of educators, business leaders, and reporters, "I don't think [voluntary national tests] will ever come about" (Hoff, 1998).

In March 1998, NAGB adopted specifications for a 4th grade reading test and an 8th grade mathematics test to be scored according to Basic, Proficient, and Advanced levels of achievement. The proposed reading test, should it come to pass, would be administered in two 45-minute segments and include multiple-choice and open-ended responses related to reading literature (such as classic and contemporary short stories, essays, biographies) and reading for information (encyclopedias, magazine articles). The proposed mathematics test,

again, should it come to pass, would also be administered in two 45-minute segments and encompass problems on number sense, measurement, geometric and spatial sense, data analysis, statistics, probability, algebra and functions. Calculators would be used for about a third of the assessment.

The American Institutes for Research, working collaboratively with several test publishers, is still developing test items for the VNTs under a $45 million contract to the U.S. Department of Education. Since the use of federal funding to pilot or field test them has, however, been banned, it is unlikely that the voluntary national testing program as envisioned by Clinton will come to pass. It is possible that the items developed for the VNT will be made available to states and local districts to include in their own testing programs (Barton, 1999). And some states are looking for easier ways to get national test results, including the possibility of inserting, or embedding, a common set of test questions into existing state assessments. Achieve, a nonpartisan nonprofit group created by state governors and business leaders, is coordinating this effort.

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

While teachers are not likely to see national tests along the lines of the ones envisioned by the Clinton administration, it's possible that the state version of the National Assessment of Educational Progress will some day perform a similar function–that is, provide comparable achievement data for students right down to the building level, if not the individual student level.

The National Assessment of Educational Progress, nicknamed "the nation's report card," is a congressionally mandated project of the National Center for Education Statistics (NCES) within the U.S. Department of Education. The NAEP assessment is administered annually by NCES to a nationally representative sample of public and private school students in grades 4, 8, and 12 to get a picture of what American children know and can do. Since its initial administration in 1969, the NAEP format has changed over time to reflect changes in testing and instructional practices. For example, NAEP was once entirely multiple choice but now includes open-ended responses. Students prepare writing samples, work with science kits, use calculators and other tools, and prepare art projects as part of the various subject-area assessments. In this respect, it is a very innovative assessment and one that has served as a model for some of the more sophisticated state testing programs.

To help states measure students' academic performance over time and to allow for cross-state comparisons, a state component was added to NAEP in 1990. Now, states can choose to administer NAEP to representative state samples in grades 4 and 8 and receive results reported by subgroups such as student gender, race/ethnicity, and parents' educational level. While participation in the state NAEP and the main NAEP are voluntary, in reality, compliance is quite high. In 2000, for example, 47 states and jurisdictions participated in the state component. This does not replace participation in the main NAEP. At present, individual scores are not gathered or reported, but the state NAEP has the potential to be used that way. Already, some districts choose to opt in to NAEP separate from the state component.

The national sample typically involves 100,000 students from 2,000 schools; state samples typically include 2,500 students per subject, per grade, drawn from 100 schools in each participating state (NCES, 1999). A key feature to keep in mind is that NAEP results are analyzed by groups rather than individual students. The names of participating schools and students are kept confidential; individual scores are not kept or released.

Two subject areas are typically assessed each year. Reading, mathematics, writing, and science are assessed most frequently, usually at 4-year intervals so that trends can be monitored. Civics, U.S. history, geography, and the arts have also been assessed in recent years, and foreign language will be assessed for the first time in 2003.

Students in participating schools are randomly selected to take one portion of the assessment being administered in a given year (usually administered during a 1-1/2 to 2-hour testing period). Achievement is reported at one of three levels: Basic, for partial mastery; Proficient, for solid academic performance, and Advanced, for superior work. A forth level, Below Basic, indicates less than acceptable performance. Again, only group and subgroup scores are reported; they are not linked back to individual students or teachers. In order to gain information about what factors correlate with student achievement, students, teachers and principals at schools participating in NAEP are also asked to complete questionnaires that address such practices as the amount of homework teachers assign and the amount of television students view. NAEP results are usually watched closely because the assessment is considered a highly respected, technically sound longitudinal measure of U.S. student achievement.

A 26-member independent board called the National Assessment Governing Board (NAGB) is responsible for setting NAEP policy, selecting which subject areas will be assessed, and overseeing the content and design of each NAEP assessment. Members include college professors, teachers, principals, superintendents, state education officials, governors, and business representatives. NAGB does not attempt to specify a national curriculum, but rather, outlines what a national assessment should test, based on a national consensus process that involves gathering input from teachers, curriculum experts, policymakers, the business community, and the public. Three contractors currently work directly on NAEP: the Educational Testing Service designs the instruments and conducts data analysis and reporting; Westat performs sampling and data collection activities; and National Computer Systems distributes materials and scores the assessments. The government also contracts for periodic research and validity studies on NAEP.

TESTS, TESTS EVERYWHERE

While almost every state has implemented some sort of state testing program, the differences in what they measure, how they measure it, and how they set achievement levels make it virtually impossible to conduct meaningful state-by-state comparisons of individual student performance. Some people believe state-to-state comparisons are irrelevant because education is a state and local function. Others believe cross-state comparisons will help spur reform and ensure uniformly high-quality education across the country. Theoretically, a state-level NAEP would yield useful data. In reality, however, NAEP state-level results have sometimes been confusing because achievement levels of students appear

to be much lower on NAEP than on the state tests. This discrepancy may be attributed to a number of factors, including the following:

- State tests are more likely to be aligned with state curricula than NAEP is.
- State tests and NAEP use different definitions of proficiency.
- State tests and NAEP may use different formats.
- State tests and NAEP differ in terms of who takes them (e.g., whether students in special education or with limited English proficiency are included).

In general, fewer students are judged to reach the Proficient standard on the NAEP reading and math tests than on state tests (GAO, 1998). This discrepancy can lead people who are not aware of the differences in the two types of tests to question the validity of their own state testing programs or the desirability of participating in a federal one.

Cost is potentially an additional barrier to nationwide testing of individual students. During the voluntary national testing debates, the General Accounting Office (1998) estimated that the per-administration cost of each test would be $12. If the assessments were administered to each of the nation's public and private school children in grades 4 and 8, the total cost would be up to $96 million, and it is not clear who would pay. Most states are already heavily invested in their own state testing programs.

It is difficult to predict how the national testing issue will ultimately be resolved. As state testing programs become institutionalized, and the public continues to be urged make judgments about school quality based on test scores, there will likely be a real push to compare results across states. Therefore, it makes sense for teachers to stay active in the discussion.

BECOMING INVOLVED IN THE NATIONAL TESTING DEBATE

- Find out whether your state is involved in the NAEP assessment program. (See http://nces.ed.gov/nationsreportcard which includes a state map with summary information and contact people.)

- Visit the NAEP Web site at http://nces.ed.gov/nationsreportcard to see sample questions, answers, frameworks, and classroom exercises in your subject area. How are these items related to your curriculum, instruction, and assessment practices?

- Take a look at the specifications for the voluntary national tests in 4th grade reading and 8th grade mathematics at www.nagb.org and follow the debate on national testing by monitoring the U.S. Department of Education's web site at www.ed.gov after the 2000 presidential election.

- Speak out. Teachers offer a valuable front-line perspective on testing. You can let your legislators know your views on the voluntary national tests through a letter or e-mail. Get addresses at http://www.congress.org or call the Capitol switchboard at (202) 224-3121

.

References

Rudner, L. and W. Schafer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.

Barton, P. E. Too much testing of the wrong kind; too little of the right kind in K-12 education. A policy information perspective. Princeton, NJ: Educational Testing Service. ED 430 052.

Davey, L. (1992). The case for a national test. *Practical Assessment, Research & Evaluation*, 3 (1). [Available online: http://ericae.net/pare/getvn.asp?v=3&n=1]

Davey, L. and Neill, M. (1992). The case against a national test. ERIC Digest. Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation. American Institutes of Research. ED 338 703.

General Accounting Office (1998). Student testing: Issues related to voluntary national mathematics and reading tests. Report to the Honorable William F. Goodling, Chairman, Committee on Education and the Workforce, House of Representatives, and the Honorable John Ashcroft, U.S. Senate. Washington, DC: Author. ED 423 244.

Hoff, D. J. (1998). Strong words underscore national testing question. Education Week, February 18, 1998.

Houston Chronicle Interactive. (1997). Debate on education focuses on national testing program. September 9, 1997.

Los Angeles Times (1997). Lott assails Clinton's plan for school tests. November 2, 1997.

National Center for Education Statistics (November 1999). The NAEP guide: A description of the content and methods of the 1999 and 2000 assessments. Washington, DC: U.S. Department of Education. [Available online at http://www.nces.ed.gov/nationsreportcard/guide/2000456.shtml].

Essential Concepts for Classroom Assessment

Writing Multiple-Choice Test Items[1]

A notable concern of many teachers is that they frequently have the task of constructing tests but have relatively little training or information to rely on in this task. The objective of this article is to set out some conventional wisdom for the construction of multiple-choice tests, which are one of the most common forms of teacher-constructed tests. The comments which follow are applicable mainly to multiple-choice tests covering fairly broad topic areas.

Before proceeding, it will be useful to establish our terms for discussing multiple-choice items. The *stem* is the introductory question or incomplete statement at the beginning of each item and this is followed by the options. The *options* consist of the answer -- the correct option -- and *distractors*--the incorrect but (we hope) tempting options.

GENERAL OBJECTIVES

As a rule, one is concerned with writing stems that are clear and parsimonious, answers that are unequivocal and chosen by the students who do best on the test, and distractors that are plausible competitors of the answer as evidenced by the frequency with which they are chosen. Lastly, and probably most important, we should adopt the attitude that items need to be developed over time in the light of evidence that can be obtained from the statistical output typically provided by a measurement services office (where tests are machine-scored) and from "expert" editorial review.

PLANNING

The primary objective in planning a test is to outline the actual course content that the test will cover. A convenient way of accomplishing this is to take 10 minutes following each class to list on an index card the important concepts covered in class and in assigned reading for that day. These cards can then be used later as a source of test items. An even more conscientious approach, of course, would be to construct the test items themselves after each class. The advantage of either of these approaches is that the resulting test is likely to be a better representation of course activity than if the test were constructed before the course or after the course, when we usually have only a fond memory or optimistic syllabus to draw from. When we are satisfied that we have an accurate description of the content areas, then all that remains is to construct items that represent specific content areas. In developing good multiple-choice items, three tasks need to be considered: writing stems, writing options, and ongoing item development. The first two are discussed in this article.

WRITING STEMS

We will first describe some basic rules for the construction of multiple-choice stems, because they are typically, though not necessarily, written before the options.

---

[1] Written by Jerard Kehoe

Reprinted with permission from A. J. Nitko and S. M. Brookhart (2007) *Educational Assessment of Students* (5th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.
Reprinted with permission from A. M. Schafer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.
From the free on-line version. To order print copies call 800 229-4200

1. Before writing the stem, identify the one point to be tested by that item. In general, the stem should not pose more than one problem, although the solution to that problem may require more than one step.

2. Construct the stem to be either an incomplete statement or a direct question, avoiding stereotyped phraseology, as rote responses are usually based on verbal stereotypes. For example, the following stems (with answers in parentheses) illustrate undesirable phraseology:

What is the biological theory of recapitulation? (Ontogeny repeats phylogeny)
Who was the chief spokesman for the "American System?" (Henry Clay)

Correctly answering these questions likely depends less on understanding than on recognizing familiar phraseology.

3. Avoid including nonfunctional words that do not contribute to the basis for choosing among the options. Often an introductory statement is included to enhance the appropriateness or significance of an item but does not affect the meaning of the problem in the item. Generally, such superfluous phrases should be excluded. For example, consider:

The American flag has three colors. One of them is (1) red (2) green (3) black
versus
One of the colors of the American flag is (1) red (2) green (3) black

In particular, irrelevant material should not be used to make the answer less obvious. This tends to place too much importance on reading comprehension as a determiner of the correct option.

4. Include as much information in the stem and as little in the options as possible. For example, if the point of an item is to associate a term with its definition, the preferred format would be to present the definition in the stem and several terms as options rather than to present the term in the stem and several definitions as options.

5. Restrict the use of negatives in the stem. Negatives in the stem usually require that the answer be a false statement. Because students are likely in the habit of searching for true statements, this may introduce an unwanted bias.

6. Avoid irrelevant clues to the correct option. Grammatical construction, for example, may lead students to reject options which are grammatically incorrect as the stem is stated. Perhaps more common and subtle, though, is the problem of common elements in the stem and in the answer. Consider the following item:

What led to the formation of the States' Rights Party?
a. The level of federal taxation
b. The demand of states for the right to make their own laws
c. The industrialization of the South
d. The corruption of federal legislators on the issue of state taxation

One does not need to know U.S. history in order to be attracted to the answer, b.

Other rules that we might list are generally commonsensical, including recommendations for independent and important items and prohibitions against complex, imprecise wording.

WRITING OPTIONS

Following the construction of the item stem, the likely more difficult task of generating options presents itself. The rules we list below are not likely to simplify this task as much as they are intended to guide our creative efforts.

1. Be satisfied with three or four well constructed options. Generally, the minimal improvement to the item due to that hard-to-come-by fifth option is not worth the effort to construct it. Indeed, all else the same, a test of 10 items each with four options is likely a better test than a test with nine items of five options each.

2. Construct distractors that are comparable in length, complexity and grammatical form to the answer, avoiding the use of such words as "always," "never," and "all." Adherence to this rule avoids some of the more common sources of biased cueing. For example, we sometimes find ourselves increasing the length and specificity of the answer (relative to distractors) in order to insure its truthfulness. This, however, becomes an easy-to-spot clue for the testwise student. Related to this issue is the question of whether or not test writers should take advantage of these types of cues to construct more tempting distractors. Surely not! The number of students choosing a distractor should depend only on deficits in the content area which the item targets and should not depend on cue biases or reading comprehension differences in "favor" of the distractor.

Stem Checklist
- **U** One point per item
- **U** Doesn't encourage rote response
- **U** Simple Wording
- **U** Short Options

Options Checklist
- **U** 3 or 4 good options
- **U** Each distractor is the same length, complexity and grammatical form
- **U** No "All of the above"
- **U** Location of correct option varies

3. Options which read "none of the above," "both a. and e. above," "all of the above," _etc_., should be avoided when the students have been instructed to choose "the best answer," which implies that the options vary in degree of correctness. On the other hand, "none of the above" is acceptable if the question is factual and is probably desirable if computation yields the answer. "All of the above" is never desirable, as one recognized distractor eliminates it and two recognized answers identify it.

4. After the options are written, vary the location of the answer on as random a basis as possible. A convenient method is to flip two (or three) coins at a time where each possible Head-Tail combination is associated with a particular location for the answer. Furthermore, if the test writer is conscientious enough to randomize the answer locations, students should be informed that the locations are randomized. (Testwise students know that for some instructors the first option is rarely the answer.)

5. If possible, have a colleague with expertise in the content area of the exam review the items for possible ambiguities, redundancies or other structural difficulties. Having completed the items we are typically so relieved that we may be tempted to regard the task as completed and each item in its final and permanent form. Yet, another source of item and test improvement is available to us, namely, statistical analyses of student responses.

This article was adapted with from *Testing Memo 4: Constructing Multiple-Choice Tests -- Part I,* Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060

Further Reading

Airasian, P. (1994) *Classroom Assessment*, Second Edition, NY: McGraw-Hill.

Cangelosi, J. (1990) *Designing Tests for Evaluating Student Achievement.* NY: Addison Wellesley.

Grunlund, N (1993) *How to make achievement tests and assessments*, 5th edition, NY: Allen and Bacon.

Haladyna, T.M. & Downing, S.M. (1989) Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, 2 (1), 51-78.

More Multiple-choice Item Writing Do's And Don'ts [1]

The previous chapter gave a few suggestions for item-writing, but only to a limited extent, due to its coverage of other aspects of test development. What follows here is a fairly comprehensive list of recommendations for writing multiple choice items. Some of these are backed up by psychometric research; i.e., it has been found that, generally, the resulting scores are more accurate indicators of each student's knowledge when the recommendations are followed than when they are violated. Other recommendations result from logical deduction.

CONTENT

1. Do ask questions that require more than knowledge of facts. For example, a question might require selection of the best answer when all of the options contain elements of correctness. Such questions tend to be more difficult and discriminating than questions that merely ask for a fact. Justifying the "bestness" of the keyed option may be as challenging to the instructor as the item was to the students, but, after all, isn't challenging students and responding to their challenges a big part of what being a teacher is all about?

More than factual recall
No superfluous information

2. Don't offer superfluous information as an introduction to a question, for example, "*The presence and association of the male seems to have profound effects on female physiology in domestic animals. Research has shown that in cattle presence of a bull has the following effect:*" This approach probably represents an unconscious effort to continue teaching while testing and is not likely to be appreciated by the students, who would prefer direct questions and less to read. The stem just quoted could be condensed to "Research has shown that the presence of a bull has which of the following effects on cows?" (17 words versus 30).

STRUCTURE

Stem and options related

3. Don't ask a question that begins, "*Which of the following is true [or false]?"* followed by a collection of unrelated options. Each test question should focus on some specific aspect of the course. Therefore, it's OK to use items that begin, "Which of the following is true [or false] concerning X?" followed by options all pertaining to X. However, this construction should be used sparingly if there is a tendency to resort to trivial reasons for falseness or an opposite tendency to offer options that are too obviously true. A few true-false questions (in among the multiple-choice questions) may forestall these problems. The options would be: *1) True 2) False.*

---

[1] Written by Robert B. Frary

Richard A. Schafer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.
From the free on-line version. To order print copies call 800 229-4200

4. Don't use items like the following:

What is (are) the capital(s) of Bolivia?
A. La Paz B. Sucre C. Santa Cruz

1) A only 4) Both A and B
2) B only 5) All of the above
3) C only

Research on this item type has consistently shown it to be easier and less discriminating than items with distinct options. In the example above, one only needs to remember that Bolivia has two capitals to be assured of answering correctly. This problem can be alleviated by offering all possible combinations of the three basic options, namely:

1) A only, 2) B only, 3) C only, 4) A and B, 5) A and C, 6) B and C, 7) A, B, and C, 8) None of the above.

However, due to its complexity, initial use of this adaptation should be limited.

OPTIONS

5. Do ask questions with varying numbers of options. There is no psychometric advantage to having a uniform number, especially if doing so results in options that are so implausible that no one or almost no one marks them. In fact, some valid and important questions demand only two or three options, e.g., *"If drug X is administered, body temperature will probably: 1) increase, 2) stay about the same, 3) decrease.*"

6. Don't put negative options following a negative stem. Empirically (or statistically) such items may appear to perform adequately, but this is probably only because brighter students who naturally tend to get higher scores are also better able to cope with the logical complexity of a double negative.

7. Don't use "*all of the above.*" Recognition of one wrong option eliminates "all of the above," and recognition of two right options identifies it as the answer, even if the other options are completely unknown to the student. Probably some instructors use items with "all of the above" as yet another way of extending their teaching into the test (see 2 above). It just seems so good to have the students affirm, say, all of the major causes of some phenomenon. With this approach, "all of the above" is the answer to almost every item containing it, and the students soon figure this out.

8. Do ask questions with "*none of the above"* as the final option, especially if the answer requires computation. Its use makes the question harder and more discriminating, because the uncertain student cannot focus on a set of options that must contain the answer. Of course, "*none of the above*" cannot be used if the question requires selection of the best answer and should not be used following a negative stem. Also, it is important that "*none of the above*" should be the answer to a reasonable proportion of the questions containing it.

9. Don't include superfluous information in the options. The reasons given for 8 above apply. In addition, as another manifestation of the desire to teach while testing, the additional information is likely to appear on the correct answer: *1) W, 2) X, 3) Y, because ...., 4) Z.* Students are very sensitive to this tendency and take advantage of it.

10. Don't use specific determiners in distractors. Sometimes in a desperate effort to produce another, often unneeded, distractor (see 5 above), a statement is made incorrect by the inclusion of words like all or never, e.g., "*All humans have 46 chromosomes.*" Students learn to classify such statements as distractors when otherwise ignorant.

11. Don't repeat wording from the stem in the correct option. Again, an ignorant student will take advantage of this practice.

ERRORS TO AVOID

Most violations of the recommendations given thus far should not be classified as outright errors, but, instead, perhaps, as lapses of judgement. And, as almost all rules have exceptions, there are probably circumstances where some of 1-11 above would not hold. However, there are three not-too-common item-writing/test-preparation errors that represent nothing less than negligence. They are now mentioned to encourage careful preparation and proofreading of tests:

OK

U Different number of option
U "None of the above" (sometimes)

AVOID

V Typos
V Inconsistent grammar
V Overlapping distractors

Typos. These are more likely to appear in distractors than in the stem and the correct answer, which get more scrutiny from the test preparer. Students easily become aware of this tendency if it is present.

Grammatical inconsistency between stem and options. Almost always, the stem and the correct answer are grammatically consistent, but distractors, often produced as afterthoughts, may not mesh properly with the stem. Again, students quickly learn to take advantage of this foible.

Overlapping distractors. For example: *Due to budget cutbacks, the university library now subscribes to fewer than _?_ periodicals. 1) 25,000 2) 20,000 3) 15,000 4) 10,000*

Perhaps surprisingly, not all students "catch on" to items like this, but many do. Worse yet, the instructor might indicate option 2 as the correct answer.

Finally, we consider an item-writing foible reported by Smith (1982). What option would you select among the following (stem omitted)?

1) Abraham Lincoln 3) Stephen A. Douglas
2) Robert E. Lee 4) Andrew Jackson

The testwise but ignorant student will select Lincoln because it represents the intersection of two categories of prominent nineteenth century people, namely, presidents and men associated with the Civil War.

Try this one:

1) before breakfast 3) on a full stomach
2) with meals 4) before going to bed

Three options have to do with eating, and two with the time of day. Only one relates to both. Unfortunately, some item writers consciously or unconsciously construct items of this type with the intersection invariably the correct answer.

This article was adapted from *Testing Memo 10: Some Multiple-choice Item Writing Do's And Don'ts*, Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060

Further Reading

Airasian, P. (1994). *Classroom Assessment*, Second Edition, NY: McGraw-Hill.

Brown, F. (1983). *Principles of Educational and Psychological Testing*, Third edition, NY: Holt Rinehart, Winston. Chapter 11.

Cangelosi, J. (1990). *Designing Tests for Evaluating Student Achievement*. NY: Longman.

Grunlund, N (1993). *How to make achievement tests and assessments*, 5th edition, NY: Allen and Bacon.

Haladyna, T.M. & Downing, S.M. (1989). Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, 2 (1), 51-78.

Kehoe, J (1995). Writing Multiple-Choice Test Items. *Practical Assessment, Research and Evaluation*, 4(4). [Available online http://ericae.net/pare/getvn.asp?v4&n4].

Roid, G.H. & Haladyna, T.M. (1980). The emergence of an item writing technology. *Review of Educational Research*, 49, 252-279.

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple-choice items. Journal of Educational Measurement, 19, 211-220.

Wesman, A.G. (1971). Writing the test item. In R.L. Thorndike (Ed.) *Educational Measurement* (1st ed, pp 99-111). Washington, DC: American Council on Education

Implementing Performance Assessment in the Classroom[1]

If you are like most teachers, it probably is a common practice for you to devise some sort of test to determine whether a previously taught concept has been learned before introducing something new to your students. Probably, this will be either a completion or multiple choice test. However, it is difficult to write completion or multiple choice tests that go beyond the recall level. For example, the results of an English test may indicate that a student knows each story has a beginning, a middle, and an end. However, these results do not guarantee that a student will write a story with a clear beginning, middle, and end. Because of this, educators have advocated the use of performance-based assessments.

Performance-based assessments "represent a set of strategies for the . . . application of knowledge, skills, and work habits through the performance of tasks that are meaningful and engaging to students" (Hibbard and others, 1996, p. 5). This type of assessment provides teachers with information about how a child understands and applies knowledge. Also, teachers can integrate performance-based assessments into the instructional process to provide additional learning experiences for students.

The benefit of performance-based assessments are well documented. However, some teachers are hesitant to implement them in their classrooms. Commonly, this is because these teachers feel they don't know enough about how to fairly assess a student's performance (Airasian,1991). Another reason for reluctance in using performance-based assessments may be previous experiences with them when the execution was unsuccessful or the results were inconclusive (Stiggins, 1994). The purpose of this chapter is to outline the basic steps that you can take to plan and execute effective performance-based assessments.

DEFINING THE PURPOSE OF THE PERFORMANCE-BASED ASSESSMENT

In order to administer any good assessment, you must have a clearly defined purpose. Thus, you must ask yourself several important questions:

- What concept, skill, or knowledge am I trying to assess?
- What should my students know?
- At what level should my students be performing?
- What type of knowledge is being assessed: reasoning, memory, or process (Stiggins, 1994)?

Ask yourself

- What am I trying to assess?
- What should the students know?
- What level?
- What type of knowledge?

[1] Written by Amy Brualdi

By answering these questions, you can decide what type of activity best suits you assessment needs.

CHOOSING THE ACTIVITY

After you define the purpose of the assessment, you can make decisions concerning the activity. There are some things that you must take into account before you choose the activity: time constraints, availability of resources in the classroom, and how much data is necessary in order to make an informed decision about the quality of a student's performance (This consideration is frequently referred to as sampling.).

The literature distinguishes between two types of performance-based assessment activities that you can implement in your classroom: informal and formal (Airasian, 1991; Popham, 1995; Stiggins, 1994). When a student is being informally assessed, the student does not know that the assessment is taking place. As a teacher, you probably use informal performance assessments all the time. One example of something that you may assess in this manner is how children interact with other children (Stiggins, 1994). You also may use informal assessment to assess a student's typical behavior or work habits.

A student who is being formally assessed knows that you are evaluating him/her. When a student's performance is formally assessed, you may either have the student perform a task or complete a project. You can either observe the student as he/she performs specific tasks or evaluate the quality of finished products.

You must beware that not all hands-on activities can be used as performance-based assessments (Wiggins, 1993). Performance-based assessments require individuals to apply their knowledge and skills in context, not merely completing a task on cue.

DEFINING THE CRITERIA

After you have determined the activity as well as what tasks will be included in the activity, you need to define which elements of the project/task you shall to determine the success of the student's performance. Sometimes, you may be able to find these criteria in local and state curriculums or other published documents (Airasian, 1991). Although these resources may prove to be very useful to you, please note that some lists of criteria may include too many skills or concepts or may not fit your needs exactly. With this in mind, you must be certain to review criteria lists before applying any of them to your performance-based assessment.

You must develop your own criteria most of the time. When you need to do this, Airasian (1991, p. 244) suggests that you complete the following steps:

- Identify the overall performance or task to be assessed, and perform it yourself or imagine yourself performing it
- List the important aspects of the performance or product.
- Try to limit the number of performance criteria, so they can all be observed during a pupil's performance.

- If possible, have groups of teachers think through the important behaviors included in a task.
- Express the performance criteria in terms of observable pupil behaviors or product characteristics.
- Don't use ambiguous words that cloud the meaning of the performance criteria.
- Arrange the performance criteria in the order in which they are likely to be observed.

Defining Criteria

1. Identify task
2. List all important aspects
3. Reduce list to fit time frame
4. Check with other teachers
5. Express as observable
6. Arrange

You may even wish to allow your students to participate in this process. You can do this by asking the students to name the elements of the project/task that they would use to determine how successfully it was completed (Stix, 1997).

Having clearly defined criteria will make it easier for you to remain objective during the assessment. The reason for this is the fact that you will know exactly which skills and/or concepts that you are supposed to be assessing. If your students were not already involved in the process of determining the criteria, you will usually want to share them with your students. This will help students know exactly what is expected of them.

CREATING PERFORMANCE RUBRICS

As opposed to most traditional forms of testing, performance-based assessments don't have clear-cut right or wrong answers. Rather, there are degrees to which a person is successful or unsuccessful. Thus, you need to evaluate the performance in a way that will allow you take those varying degrees into consideration. This can be accomplished by creating rubrics.

A rubric is a rating system by which teachers can determine at what level of proficiency a student is able to perform a task or display knowledge of a concept. With rubrics, you can define the different levels of proficiency for each criterion. Like the process of developing criteria, you can either utilize previously developed rubrics or create your own. When using any type of rubric, you need to be certain that the rubrics are fair and simple. Also, the performance at each level must be clearly defined and accurately reflect its corresponding criterion (or subcategory) (Airasian, 1991; Popham, 1995; Stiggins, 1994).

When deciding how to communicate the varying levels of proficiency, you may wish to use impartial words instead of numerical or letter grades (Stix, 1997). For instance, you may want to use the following scale: word, sentence, page, chapter, book. However, words such as "novice," "apprentice," "proficient," and "excellent" are frequently used.

As with criteria development, allowing your students to assist in the creation of rubrics may be a good learning experience for them. You can engage students in this process by showing them examples of the same task performed/project completed at different levels and discuss to what degree the different elements of the criteria were displayed. However, if your students do not help to create the different rubrics, you will probably want to share those rubrics with your students before they complete the task or project.

ASSESSING THE PERFORMANCE

Using this information, you can give feedback on a student's performance either in the form of a narrative report or a grade. There are several different ways to record the results of performance-based assessments (Airasian,1991; Stiggins,1994):

- Checklist Approach When you use this, you only have to indicate whether or not certain elements are present in the performances.
- Narrative/Anecdotal Approach When teachers use this, they will write narrative reports of what was done during each of the performances. From these reports, teachers can determine how well their students met their standards.
- Rating Scale Approach When teachers use this, they indicate to what degree the standards were met. Usually, teachers will use a numerical scale. For instance, one teacher may rate each criterion on a scale of one to five with one meaning "skill barely present" and five meaning "skill extremely well executed."
- Memory Approach When teachers use this, they observe the students performing the tasks without taking any notes. They use the information from their memory to determine whether or not the students were successful. (Please note that this approach is not recommended.)

While it is a standard procedure for teachers to assess students' performances, teachers may wish to allow students to assess them themselves. Permitting students to do this provides them with the opportunity to reflect upon the quality of their work and learn from their successes and failures.

References and Additional Reading

Airasian, P.W. (1991). *Classroom assessment*. New York : McGraw-Hill.

Hibbard, K. M. and others. (1996). *A teacher's guide to performance-based learning and assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J. (1995). *Classroom assessment: What teachers need to know*. Needham Heights, MA: Allyn and Bacon.

Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York: Macmillan Publishing Company.

Stix, A. (1997). Empowering students through negotiable contracting. (Paper presented at the National Middle School Initiative Conference (Long Island, NY, January 25, 1997) (ERIC Document Reproduction Number ED411274)

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, May, 703-713.
Wiggins, G. (1993). Assessment, authenticity, context, and validity. *Phi Delta Kappan*, November, 200-214.
Wiggins,G. (1998). *Educative assessment: designing assessments to inform and improve student performance* San Francisco, Calif. : Jossey-Bass.

Scoring Rubrics: What, When and How?[1]

Scoring rubrics have become a common method for evaluating student work in both the K-12 and the college classrooms. The purpose of this paper is to describe the different types of scoring rubrics, explain why scoring rubrics are useful and provide a process for developing scoring rubrics. This paper concludes with a description of resources that contain examples of the different types of scoring rubrics and further guidance in the development process.

WHAT IS A SCORING RUBRIC?

Scoring rubrics are descriptive scoring schemes that are developed by teachers or other evaluators to guide the analysis of the products or processes of students' efforts (Brookhart, 1999). Scoring rubrics are typically employed when a judgement of quality is required and may be used to evaluate a broad range of subjects and activities. One common use of scoring rubrics is to guide the evaluation of writing samples. Judgements concerning the quality of a given writing sample may vary depending upon the criteria established by the individual evaluator. One evaluator may heavily weigh the evaluation process upon the linguistic structure, while another evaluator may be more interested in the persuasiveness of the argument. A high quality essay is likely to have a combination of these and other factors. By developing a pre-defined scheme for the evaluation process, the subjectivity involved in evaluating an essay becomes more objective.

Figure 1 displays a scoring rubric that was developed to guide the evaluation of student writing samples in a college classroom (based loosely on Leydens & Thompson, 1997). This is an example of a holistic scoring rubric with four score levels. Holistic rubrics will be discussed in detail later in this document. As the example illustrates, each score category describes the characteristics of a response that would receive the respective score. By having a description of the characteristics of responses within each score category, the likelihood that two independent evaluators would assign the same score to a given response is increased. This concept of examining the extent to which two independent evaluators assign the same score to a given response is referred to as "rater reliability."

Figure 1.
Example of a scoring rubric designed to evaluate college writing samples.

-3-
Meets Expectations for a first Draft of a Professional Report

g The document can be easily followed. A combination of the following are apparent in the document:

[1] Written by Barbara M. Moskal.

1. Effective transitions are used throughout,
2. A professional format is used,
3. The graphics are descriptive and clearly support the document''s purpose.

g The document is clear and concise and appropriate grammar is used throughout.

-2-
Adequate

g The document can be easily followed. A combination of the following are apparent in the document:

1. Basic transitions are used,
2. A structured format is used,
3. Some supporting graphics are provided, but are not clearly explained.

g The document contains minimal distractions that appear in a combination of the following forms:

1. Flow in thought
2. Graphical presentations
3. Grammar/mechanics

-1-
Needs Improvement

g Organization of document is difficult to follow due to a combination of following:

1. Inadequate transitions
2. Rambling format
3. Insufficient or irrelevant information
4. Ambiguous graphics

g The document contains numerous distractions that appear in the a combination of the following forms:

1. Flow in thought
2. Graphical presentations
3. Grammar/mechanics

-0-
Inadequate

- There appears to be no organization of the document''s contents.
- Sentences are difficult to read and understand.

WHEN ARE SCORING RUBRICS AN APPROPRIATE EVALUATION TECHNIQUE?

Writing samples are just one example of performances that may be evaluated using scoring rubrics. Scoring rubrics have also been used to evaluate group activities, extended projects and oral presentations (e.g., Chicago Public Schools, 1999; Danielson, 1997a; 1997b; Schrock, 2000; Moskal, 2000). They are equally appropriate to the English, Mathematics and Science classrooms (e.g., Chicago Public Schools, 1999; State of Colorado, 1999; Danielson, 1997a; 1997b; Danielson & Marquez, 1998; Schrock, 2000). Both pre-college and college instructors use scoring rubrics for classroom evaluation purposes (e.g., State of Colorado, 1999; Schrock, 2000; Moskal, 2000; Knecht, Moskal & Pavelich, 2000). Where and when a scoring rubric is used does not depend on the grade level or subject, but rather on the purpose of the assessment.

Scoring rubrics are one of many alternatives available for evaluating student work. For example, checklists may be used rather then scoring rubrics in the evaluation of writing samples. Checklists are an appropriate choice for evaluation when the information that is sought is limited to the determination of whether specific criteria have been met. Scoring rubrics are based on descriptive scales and support the evaluation of the extent to which criteria has been met.

The assignment of numerical weights to sub-skills within a process is another evaluation technique that may be used to determine the extent to which given criteria has been met. Numerical values, however, do not provide students with an indication as to how to improve their performance. A student who receives a "70" out of "100", may not know how to improve his or her performance on the next assignment. Scoring rubrics respond to this concern by providing descriptions at each level as to what is expected. These descriptions assist the students in understanding why they received the score that they did and what they need to do to improve their future performances.

Whether a scoring rubric is an appropriate evaluation technique is dependent upon the purpose of the assessment. Scoring rubrics provide at least two benefits in the evaluation process. First, they support the examination of the extent to which the specified criteria has been reached. Second, they provide feedback to students concerning how to improve their performances. If these benefits are consistent with the purpose of the assessment, than a scoring rubric is likely to be an appropriate evaluation technique.

WHAT ARE THE DIFFERENT TYPES OF SCORING RUBRICS?

Several different types of scoring rubrics are available. Which variation of the scoring rubric should be used in a given evaluation is also dependent upon the purpose of the evaluation. This section describes the differences between analytic and holistic scoring rubrics and between task specific and general scoring rubrics.

Analytic verses Holistic

In the initial phases of developing a scoring rubric, the evaluator needs to determine what will be the evaluation criteria. For example, two factors that may be considered in the evaluation of a writing sample are whether appropriate grammar is used and the extent to which the given argument is persuasive. An analytic scoring rubric, much like the checklist, allows for the separate evaluation of each of these factors. Each criterion is scored on a different descriptive scale (Brookhart, 1999).

The rubric that is displayed in Figure 1 could be extended to include a separate set of criteria for the evaluation of the persuasiveness of the argument. This extension would result in an analytic scoring rubric with two factors, quality of written expression and persuasiveness of the argument. Each factor would receive a separate score. Occasionally, numerical weights are assigned to the evaluation of each criterion. As discussed earlier, the benefit of using a scoring rubric rather than weighted scores is that scoring rubrics provide a description of what is expected at each score level. Students may use this information to improve their future performance.

Occasionally, it is not possible to separate an evaluation into independent factors. When there is an overlap between the criteria set for the evaluation of the different factors, a holistic scoring rubric may be preferable to an analytic scoring rubric. In a holistic scoring rubric, the criteria is considered in combination on a single descriptive scale (Brookhart, 1999). Holistic scoring rubrics support broader judgements concerning the quality of the process or the product.

Selecting to use an analytic scoring rubric does not eliminate the possibility of a holistic factor. A holistic judgement may be built into an analytic scoring rubric as one of the score categories. One difficulty with this approach is that overlap between the criteria that is set for the holistic judgement and the other evaluated factors cannot be avoided. When one of the purposes of the evaluation is to assign a grade, this overlap should be carefully considered and controlled. The evaluator should determine whether the overlap is resulting in certain criteria are being weighted more than was originally intended. In other words, the evaluator needs to be careful that the student is not unintentionally severely penalized for a given mistake.

General verses Task Specific

Scoring rubrics may be designed for the evaluation of a specific task or

Use descriptors rather than judgements.

theevaluation of a broader category of tasks. If the purpose of a given course is to develop a student's oral communication skills, a general scoring rubric may be developed and used to evaluate each of the oral presentations given by that student. This approach would allow the students to use the feedback that they acquired from the last presentation to improve their performance on the next presentation.

If each oral presentation focuses upon a different historical event and the purpose of the assessment is to evaluate the students' knowledge of the given event, a general scoring rubric for evaluating a sequence of presentations may not be adequate. Historical events differ in both influencing factors and outcomes. In order to evaluate the students' factual and conceptual knowledge of these events, it may be necessary to develop separate scoring rubrics for each presentation. A "Task Specific" scoring rubric is designed to evaluate student performances on a single assessment event.

Scoring rubrics may be designed to contain both general and task specific components. If the purpose of a presentation is to evaluate students' oral presentation skills and their knowledge of the historical event that is being discussed, an analytic rubric could be used that contains both a general component and a task specific component. The oral component of the rubric may consist of a general set of criteria developed for the evaluation of oral presentations; the task specific component of the rubric may contain a set of criteria developed with the specific historical event in mind.

HOW ARE SCORING RUBRICS DEVELOPED?

The first step in developing a scoring rubric is to clearly identify the qualities that need to be displayed in a student's work to demonstrate proficient performance (Brookhart, 1999). The identified qualities will form the top level or levels of scoring criteria for the scoring rubric. The decision can then be made as to whether the information that is desired from the evaluation can best be acquired through the use of an analytic or holistic scoring rubric. If an analytic scoring rubric is created, then each criterion is considered separately as the descriptions of the different score levels are developed. This process results in separate descriptive scoring schemes for each evaluation factor. For holistic scoring rubrics, the collection of criteria is considered throughout the construction of each level of the scoring rubric and the result is a single descriptive scoring scheme.

Steps in developing a scoring rubric

1. Identify qualities for the highest score
2. Select analytic or holistic scoing
3. If analytic, develop scoring schemes for each factor
4. Define criteria for lowest level
5. Contrast lowest and highest to develop middle level
6. Contract other levels for finer distinctions

After defining the criteria for the top level of performance, the evaluator's attention may be turned to defining the criteria for lowest level of performance. What type of performance would suggest a very limited understanding of the concepts that

are being assessed? The contrast between the criteria for top level performance and bottom level performance is likely to suggest appropriate criteria for middle level of performance. This approach would result in three score levels.

If greater distinctions are desired, then comparisons can be made between the criteria for each existing score level. The contrast between levels is likely to suggest criteria that may be used to create score levels that fall between the existing score levels. This comparison process can be used until the desired number of score levels is reached or until no further distinctions can be made. If meaningful distinctions between the score categories cannot be made, then additional score categories should not be created (Brookhart, 1999). It is better to have a few meaningful score categories then to have many score categories that are difficult or impossible to distinguish.

Each score category should be defined using descriptions of the work rather then judgements about the work (Brookhart, 1999). For example, "Student's mathematical calculations contain no errors," is preferable over, "Student's calculations are good." The phrase "are good" requires the evaluator to make a judgement whereas the phrase "no errors" is quantifiable. In order to determine whether a rubric provides adequate descriptions, another teacher may be asked to use the scoring rubric to evaluate a sub-set of student responses. Differences between the scores assigned by the original rubric developer and the second scorer will suggest how the rubric may be further clarified.

RESOURCES

Currently, there is a broad range of resources available to teachers who wish to use scoring rubrics in their classrooms. These resources differ both in the subject that they cover and the level that they are designed to assess. The examples provided below are only a small sample of the information that is available.

For K-12 teachers, the State of Colorado (1998) has developed an on-line set of general, holistic scoring rubrics that are designed for the evaluation of various writing assessments. The Chicago Public Schools (1999) maintain an extensive electronic list of analytic and holistic scoring rubrics that span the broad array of subjects represented throughout K-12 education. For mathematics teachers, Danielson has developed a collection of reference books that contain scoring rubrics that are appropriate to the elementary, middle school and high school mathematics classrooms (1997a, 1997b; Danielson & Marquez, 1998).

Resources are also available to assist college instructors who are interested in developing and using scoring rubrics in their classrooms. *Kathy Schrock's Guide for Educators* (2000) contains electronic materials for both the pre-college and the college classroom. In *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*, Brookhart (1999) provides a brief, but comprehensive review of the literature on assessment in the college classroom. This includes a description of scoring rubrics and why their use is increasing in the college classroom. Moskal (1999) has developed a web site that contains links to a variety of college assessment resources, including scoring rubrics.

The resources described above represent only a fraction of those that are available. The ERIC Clearinghouse on Assessment and Evaluation [ERIC/AE] provides several additional useful web sites. One of these, *Scoring Rubrics - Definitions & Constructions* (2000b), specifically addresses questions that are frequently asked with regard to scoring rubrics. This site also provides electronic links to web resources and bibliographic references to books and articles that discuss scoring rubrics. For more recent developments within assessment and evaluation, a search can be completed on the abstracts of papers that will soon be available through ERIC/AE (2000a). This site also contains a direct link to ERIC/AE abstracts that are specific to scoring rubrics.

Search engines that are available on the web may be used to locate additional electronic resources. When using this approach, the search criteria should be as specific as possible. Generic searches that use the terms "rubrics" or "scoring rubrics" will yield a large volume of references. When seeking information on scoring rubrics from the web, it is advisable to use an advanced search and specify the grade level, subject area and topic of interest. If more resources are desired than result from this conservative approach, the search criteria can be expanded.

References

Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27, No.1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Chicago Public Schools (1999). *Rubric Bank*. [Available online at: http://intranet.cps.k12.il.us/Assessments/Ideas_and_Rubrics/Rubric_Bank/rubric_bank.html].

Danielson, C. (1997a). *A Collection of Performance Tasks and Rubrics: Middle School Mathematics.* Larchmont, NY: Eye on Education Inc.

Danielson, C. (1997b). *A Collection of Performance Tasks and Rubrics: Upper Elementary School Mathematics*. Larchmont, NY: Eye on Education Inc.

Danielson, C. & Marquez, E. (1998). *A Collection of Performance Tasks and Rubrics: High School Mathematics*. Larchmont, NY: Eye on Education Inc.

ERIC/AE (2000a). *Search ERIC/AE draft abstracts*. [Available online at: http://ericae.net/sinprog.htm].

ERIC/AE (2000b). *Scoring Rubrics - Definitions & Construction* [Available online at: http://ericae.net/faqs/rubrics/scoring_rubrics.htm].

Knecht, R., Moskal, B. & Pavelich, M. (2000). *The Design Report Rubric: Measuring and Tracking Growth through Success*, Paper to be presented at the annual meeting of the American Society for Engineering Education.

Leydens, J. & Thompson, D. (August, 1997), *Writing Rubrics Design (EPICS) I*, Internal Communication, Design (EPICS) Program, Colorado School of Mines.

Moskal, B. (2000). *Assessment Resource Page*. [Available online at: http://www.mines.edu/Academic/assess/Resource.htm].

Schrock, K. (2000). *Kathy Schrock's Guide for Educators*. [Available online at: http://school.discovery.com/schrockguide/assess.html].

State of Colorado (1998). The Rubric. [Available online at: http://www.cde.state.co.us/cdedepcom/asrubric.htm#writing].

Scoring Rubric Development: Validity and Reliability[1]

In the previous chapter, a framework for developing scoring rubrics was presented and the issues of validity and reliability were given cursory attention. Although many teachers have been exposed to the statistical definitions of the terms "validity" and "reliability" in teacher preparation courses, these courses often do not discuss how these concepts are related to classroom practices (Stiggins, 1999). One purpose of this article is to provide clear definitions of the terms "validity" and "reliability" and illustrate these definitions through examples. A second purpose is to clarify how these issues may be addressed in the development of scoring rubrics. Scoring rubrics are descriptive scoring schemes that are developed by teachers or other evaluators to guide the analysis of the products and/or processes of students' efforts (Brookhart, 1999; Moskal, 2000). The ideas presented here are applicable for anyone using scoring rubrics in the classroom, regardless of the discipline or grade level.

VALIDITY

Validation is the process of accumulating evidence that supports the appropriateness of the inferences that are made of student responses for specified assessment uses. Validity refers to the degree to which the evidence supports that these interpretations are correct and that the manner in which the interpretations are used is appropriate (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). Three types of evidence are commonly examined to support the validity of an assessment instrument: content, construct, and criterion. This section begins by defining these types of evidence and is followed by a discussion of how evidence of validity should be considered in the development of scoring rubrics.

Content-Related Evidence

Content-related evidence refers to the extent to which a student's responses to a given assessment instrument reflects that student's knowledge of the content area that is of interest. For example, a history exam in which the questions use complex sentence structures may unintentionally measure students' reading comprehension skills rather than their historical knowledge. A teacher who is interpreting a student's incorrect response may conclude that the student does not have the appropriate historical knowledge when actually that student does not understand the questions. The teacher has misinterpreted the evidence——rendering the interpretation invalid.

Content-related evidence is also concerned with the extent to which the assessment instrument adequately samples the content domain. A mathematics test that primarily includes addition problems would provide inadequate evidence of a student's ability to solve subtraction, multiplication and division problems. Correctly computing fifty

---

[1] Written by Barbara M. Moskal & Jon A. Leydens

Rudner, L. and W. Schafer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.
From the free on-line version. To order print copies call 800 229-4200

addition problems and two multiplication problems does not provide convincing evidence that a student can subtract, multiply or divide.

Content-related evidence should also be considered when developing scoring rubrics. The task shown in Figure 1 was developed by the Quantitative Understanding: Amplifying Student Achievement and Reasoning Project (Lane, et. al, 1995) and requests that the student provide an explanation. The intended content of this task is decimal density. In developing a scoring rubric, a teacher could unintentionally emphasize the nonmathematical components of the task. For example, the resultant scoring criteria may emphasize sentence structure and/or spelling at the expense of the mathematical knowledge that the student displays. The student's score, which is interpreted as an indicator of the student's mathematical knowledge, would actually be a reflection of the student's grammatical skills. Based on this scoring system, the resultant score would be an inaccurate measure of the student's mathematical knowledge. This discussion does not suggest that sentence structure and/or spelling cannot be assessed through this task. If the assessment is intended to examine sentence structure, spelling, *and* mathematics, then the score categories should reflect all of these areas.

Figure 1. Decimal Density Task

Dena tried to identify all the numbers between 3.4 and 3.5. Dena said, "3.41, 3.42, 3.43, 3.44, 3.45, 3.46, 3.47, 3.48 and 3.49. That's all the numbers that are between 3.4 and 3.5."

Nakisha disagreed and said that there were more numbers between 3.4 and 3.5.

A. Which girl is correct?

Answer:

B. Why do you think she is correct?

Construct-Related Evidence

Constructs are processes that are internal to an individual. An example of a construct is an individual's reasoning process. Although reasoning occurs inside a person, it may be partially displayed through results and explanations. An isolated correct answer, however, does not provide clear and convincing evidence of the nature of the individual's underlying reasoning process. Although an answer results from a student's reasoning process, a correct answer may be the outcome of incorrect reasoning. When the purpose of an assessment is to evaluate reasoning, both the product (i.e., the answer) and the process (i.e., the explanation) should be requested and examined.

Consider the problem shown in Figure 1. Part A of this problem requests that the student indicate which girl is correct. Part B requests an explanation. The intention of combining these two questions into a single task is to elicit evidence of the students'

reasoning process. If a scoring rubric is used to guide the evaluation of students' responses to this task, then that rubric should contain criteria that addresses both the product and the process. An example of a holistic scoring rubric that examines both the answer and the explanation for this task is shown in Figure 2.

Figure 2. Example Rubric for Decimal Density Task

Proficient:	Answer to part A is Nakisha. Explanation clearly indicates that there are more numbers between the two given values.
Partially Proficient:	Answer to part A is Nakisha. Explanation indicates that there are a finite number of rational numbers between the two given values.
Not Proficient:	Answer to part A is Dana. Explanation indicates that all of the values between the two given values are listed.

Note. This rubric is intended as an example and was developed by the authors. It is not the original QUASAR rubric, which employs a five-point scale.

Evaluation criteria within the rubric may also be established that measure factors that are unrelated to the construct of interest. This is similar to the earlier example in which spelling errors were being examined in a mathematics assessment. However, here the concern is whether the elements of the responses being evaluated are appropriate indicators of the underlying construct. If the construct to be examined is reasoning, then spelling errors in the student's explanation are irrelevant to the purpose of the assessment and should not be included in the evaluation criteria. On the other hand, if the purpose of the assessment is to examine spelling and reasoning, then both should be reflected in the evaluation criteria. Construct-related evidence is the evidence that supports that an assessment instrument is completely and only measuring the intended construct.

Reasoning is not the only construct that may be examined through classroom assessments. Problem solving, creativity, writing process, self-esteem, and attitudes are other constructs that a teacher may wish to examine. Regardless of the construct, an effort should be made to identify the facets of the construct that may be displayed and that would provide convincing evidence of the students' underlying processes. These facets should then be carefully considered in the development of the assessment instrument and in the establishment of scoring criteria.

Criterion-Related Evidence

The final type of evidence that will be discussed here is criterion-related evidence. This type of evidence supports the extent to which the results of an assessment correlate with a current or future event. Another way to think of criterion-related evidence is to consider the extent to which the students' performance on the given task may be generalized to other, more relevant activities (Rafilson, 1991).

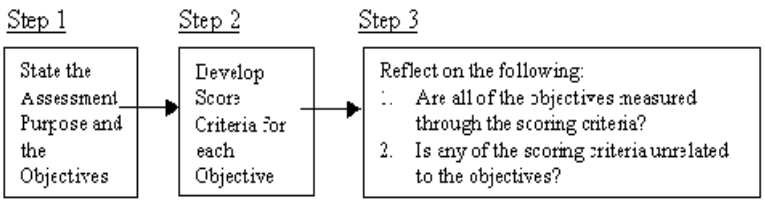
A common practice in many engineering colleges is to develop a course that "mimics" the working environment of a practicing engineer (e.g., Sheppard, & Jeninson, 1997; King, Parker, Grover, Gosink, & Middleton, 1999). These courses are specifically designed to provide the students with experiences in "real" working environments. Evaluations of these courses, which sometimes include the use of scoring rubrics (Leydens & Thompson, 1997; Knecht, Moskal & Pavelich, 2000), are intended to examine how well prepared the students are to function as professional engineers. The quality of the assessment is dependent upon identifying the components of the current environment that will suggest successful performance in the professional environment. When a scoring rubric is used to evaluate performances within these courses, the scoring criteria should address the components of the assessment activity that are directly related to practices in the field. In other words, high scores on the assessment activity should suggest high performance outside the classroom or at the future work place.

Validity Concerns in Rubric Development

Concerns about the valid interpretation of assessment results should begin before the selection or development of a task or an assessment instrument. A well-designed scoring rubric cannot correct for a poorly designed assessment instrument. Since establishing validity is dependent on the purpose of the assessment, teachers should clearly state what they hope to learn about the responding students (i.e., the purpose) and how the students will display these proficiencies (i.e., the objectives). The teacher should use the stated purpose and objectives to guide the development of the scoring rubric.

In order to ensure that an assessment instrument elicits evidence that is appropriate to the desired purpose, Hanny (2000) recommended numbering the intended objectives of a given assessment and then writing the number of the appropriate objective next to the question that addresses that objective. In this manner, any objectives that have not been addressed through the assessment will become apparent. This method for examining an assessment instrument may be modified to evaluate the appropriateness of a scoring rubric. First, clearly state the purpose and objectives of the assessment. Next, develop scoring criteria that address each objective. If one of the objectives is not represented in the score categories, then the rubric is unlikely to provide the evidence necessary to examine the given objective. If some of the scoring criteria are not related to the objectives, then, once again, the appropriateness of the assessment and the rubric is in question. This process for developing a scoring rubric is illustrated in Figure 3.

Figure 3. Evaluating the Appropriateness of Scoring Categories to a Stated Purpose



Ru

Assessment. Washington, DC: National Education Association.

Reflecting on the purpose and the objectives of the assessment will also suggest which forms of evidence——content, construct, and/or criterion——should be given consideration. If the intention of an assessment instrument is to elicit evidence of an individual's knowledge within a given content area, such as historical facts, then the appropriateness of the content-related evidence should be considered. If the assessment instrument is designed to measure reasoning, problem solving or other processes that are internal to the individual and, therefore, require more indirect examination, then the appropriateness of the construct-related evidence should be examined. If the purpose of the assessment instrument is to elicit evidence of how a student will perform outside of school or in a different situation, criterion-related evidence should be considered.

Being aware of the different types of evidence that support validity throughout the rubric development process is likely to improve the appropriateness of the interpretations when the scoring rubric is used. Validity evidence may also be examined after a preliminary rubric has been established. Table 1 displays a list of questions that may be useful in evaluating the appropriateness of a given scoring rubric with respect to the stated purpose. This table is divided according to the type of evidence being considered.

Many assessments serve multiple purposes. For example, the problem displayed in Figure 1 was designed to measure both students' knowledge of decimal density and the reasoning process that students used to solve the problem. When multiple purposes are served by a given assessment, more than one form of evidence may need to be considered.

Another form of validity evidence that is often discussed is "consequential evidence". Consequential evidence refers to examining the consequences or uses of the assessment results. For example, a teacher may find that the application of the scoring rubric to the evaluation of male and female performances on a given task consistently results in lower evaluations for the male students. The interpretation of this result may be the male students are not as proficient within the area that is being investigated as the female students. It is possible that the identified difference is actually the result of a factor that is unrelated to the purpose of the assessment. In other words, the completion of the task may require knowledge of content or constructs that were not consistent with the original purposes. Consequential evidence refers to examining the outcomes of an assessment and using these outcomes to identify possible alternative interpretations of the assessment results (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999).

Table 1: Questions to Examine Each Type of Validity Evidence

Content	Construct	Criterion
Do the evaluation criteria address any extraneous content? Do the evaluation criteria of the scoring rubric address all aspects of the intended content? Is there any content addressed in the task that should be evaluated through the rubric, but is not?	Are all of the important facets of the intended construct evaluated through the scoring criteria? Is any of the evaluation criteria irrelevant to the construct of interest?	How do the scoring criteria reflect competencies that would suggest success on future or related performances? What are the important components of the future or related performance that may be evaluated through the use of the assessment instrument? How do the scoring criteria measure the important components of the future or related performance? Are there any facets of the future or related performance that are not reflected in the scoring criteria?

RELIABILITY

Reliability refers to the consistency of assessment scores. For example, on a reliable test, a student would expect to attain the same score regardless of when the student completed the assessment, when the response was scored, and who scored the response. On an unreliable examination, a student's score may vary based on factors that are not related to the purpose of the assessment.

Many teachers are probably familiar with the terms "test/retest reliability," "equivalent-forms reliability," "split half reliability" and "rational equivalence reliability" (Gay, 1987). Each of these terms refers to statistical methods that are used to establish consistency of student performances within a given test or across more than one test. These types of reliability are of more concern on standardized or high stakes testing than they are in classroom assessment. In a classroom, students' knowledge is repeatedly assessed and this allows the teacher to adjust as new insights are acquired.

The two forms of reliability that typically are considered in classroom assessment and in rubric development involve rater (or scorer) reliability. Rater reliability generally refers to the consistency of scores that are assigned by two independent raters and that are assigned by the same rater at different points in time. The former is referred to as "interrater reliability" while the latter is referred to as "intrarater reliability."

Interrater Reliability

Interrater reliability refers to the concern that a student's score may vary from rater to rater. Students often criticize exams in which their score appears to be based on the subjective judgment of their instructor. For example, one manner in which to analyze an essay exam is to read through the students' responses and make judgments as to the quality of the students' written products. Without set criteria to guide the rating process, two independent raters may not assign the same score to a given response. Each rater has his or her own evaluation criteria. Scoring rubrics respond to this concern by formalizing the criteria at each score level. The descriptions of the score levels are used to guide the evaluation process. Although scoring rubrics do not completely eliminate variations between raters, a well-designed scoring rubric can reduce the occurrence of these discrepancies.

Intrarater Reliability

Factors that are external to the purpose of the assessment can impact the manner in which a given rater scores student responses. For example, a rater may become fatigued with the scoring process and devote less attention to the analysis over time. Certain responses may receive different scores than they would have had they been scored earlier in the evaluation. A rater's mood on the given day or knowing who a respondent is may also impact the scoring process. A correct response from a failing student may be more critically analyzed than an identical response from a student who is known to perform well. Intrarater reliability refers to each of these situations in which the scoring process of a given rater changes over time. The inconsistencies in the scoring process result from influences that are internal to the rater rather than true differences in student performances. Well-designed scoring rubrics respond to the concern of intrarater reliability by establishing a description of the scoring criteria in advance. Throughout the scoring process, the rater should revisit the established criteria in order to ensure that consistency is maintained.

Reliability Concerns in Rubric Development

Clarifying the scoring rubric is likely to improve both interrater and intrarater reliability. A scoring rubric with well-defined score categories should assist in maintaining consistent scoring regardless of who the rater is or when the rating is completed. The following questions may be used to evaluate the clarity of a given rubric: 1) Are the scoring categories well defined? 2) Are the differences between the score categories clear? And 3) Would two independent raters arrive at the same score for a given response based on the scoring rubric? If the answer to any of these questions is "no", then the unclear score categories should be revised.

One method of further clarifying a scoring rubric is through the use of anchor papers. Anchor papers are a set of scored responses that illustrate the nuances of the scoring rubric. A given rater may refer to the anchor papers throughout the scoring process to illuminate the differences between the score levels.

After every effort has been made to clarify the scoring categories, other teachers may be asked to use the rubric and the anchor papers to evaluate a sample set of responses. Any discrepancies between the scores that are assigned by the teachers will suggest which components of the scoring rubric require further explanation. Any differences in interpretation should be discussed and appropriate adjustments to the scoring rubric should be negotiated. Although this negotiation process can be time consuming, it can also greatly enhance reliability (Yancey, 1999).

Another reliability concern is the appropriateness of the given scoring rubric to the population of responding students. A scoring rubric that consistently measures the performances of one set of students may not consistently measure the performances of a different set of students. For example, if a task is embedded within a context, one population of students may be familiar with that context and the other population may be unfamiliar with that context. The students who are unfamiliar with the given context may achieve a lower score based on their lack of knowledge of the context. If these same students had completed a different task that covered the same material that was embedded in a familiar context, their scores may have been higher. When the cause of variation in performance and the resulting scores is unrelated to the purpose of the assessment, the scores are unreliable.

Sometimes during the scoring process, teachers realize that they hold implicit criteria that are not stated in the scoring rubric. Whenever possible, the scoring rubric should be shared with the students in advance in order to allow students the opportunity to construct the response with the intention of providing convincing evidence that they have met the criteria. If the scoring rubric is shared with the students prior to the evaluation, students should not be held accountable for the unstated criteria. Identifying implicit criteria can help the teacher refine the scoring rubric for future assessments.

CONCLUDING REMARKS

Establishing reliability is a prerequisite for establishing validity (Gay, 1987). Although a valid assessment is by necessity reliable, the contrary is not true. A reliable assessment is not necessarily valid. A scoring rubric is likely to result in invalid interpretations, for example, when the scoring criteria are focused on an element of the response that is not related to the purpose of the assessment. The score criteria may be so well stated that any given response would receive the same score regardless of who the rater is or when the response is scored.

A final word of caution is necessary concerning the development of scoring rubrics. Scoring rubrics describe general, synthesized criteria that are witnessed across individual performances and therefore, cannot possibly account for the unique characteristics of every performance (Delandshere & Petrosky, 1998; Haswell &

Wyche-Smith, 1994). Teachers who depend solely upon the scoring criteria during the evaluation process may be less likely to recognize inconsistencies that emerge between the observed performances and the resultant score. For example, a reliable scoring rubric may be developed and used to evaluate the performances of pre-service teachers while those individuals are providing instruction. The existence of scoring criteria may shift the rater's focus from the *interpretation* of an individual teacher''s performances to the mere *recognition* of traits that appear on the rubric (Delandshere & Petrosky, 1998). A pre-service teacher who has a unique, but effective style, may acquire an invalid, low score based on the traits of the performance.

The purpose of this article was to define the concepts of validity and reliability and to explain how these concepts are related to scoring rubric development. The reader may have noticed that the different types of scoring rubrics——analytic, holistic, task specific, and general——were not discussed here (for more on these, see Moskal, 2000). Neither validity nor reliability is dependent upon the type of rubric. Carefully designed analytic, holistic, task specific, and general scoring rubrics have the potential to produce valid and reliable results.

References

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

Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27, No.1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Delandshere, G. & Petrosky, A. (1998) "Assessment of complex performances: Limitations of key measurement assumptions." *Educational Researcher, 27* (2), 14-25.

Gay, L.R. (1987). "Selection of measurement instruments." In *Educational Research: Competencies for Analysis and Application* (3rd ed.). New York: Macmillan.

Hanny, R. J. (2000). *Assessing the SOL in classrooms*. College of William and Mary. [Available online: http://www.wm.edu/education/SURN/solass.html].

Haswell, R., & Wyche-Smith, S. (1994) "Adventuring into writing assessment." *College Composition and Communication, 45,* 220-236.

King, R.H., Parker, T.E., Grover, T.P., Gosink, J.P. & Middleton, N.T. (1999). "A multidisciplinary engineering laboratory course." *Journal of Engineering Education, 88* (3) 311- 316.

Knecht, R., Moskal, B. & Pavelich, M. (2000). *The design report rubric: Measuring and tracking growth through success.* Proceedings of the Annual Meeting American Society for Engineering Education, St. Louis, Missouri.

Lane, S., Silver, E.A., Ankenmann, R.D., Cai, J., Finseth, C., Liu, M., Magone, M.E., Meel, D., Moskal, B., Parke, C.S., Stone, C.A., Wang, N., & Zhu, Y. (1995). *QUASAR Cognitive Assessment Instrument (QCAI).* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Leydens, J. & Thompson, D. (1997, August). *Writing rubrics design (EPICS) I*, Internal Communication, Design (EPICS) Program, Colorado School of Mines.

Moskal, B. M. (2000). "Scoring rubrics: What, when and how?" *Practical Assessment, Research & Evaluation, 7* (3) [Available Online: http://ericae.net/pare/getvn.asp?v=7&n=3].

Rafilson, F. (1991). "The case for validity generalization." *Practical Assessment, Research & Evaluation, 2* (13). [Available online: http://ericae.net/pare/getvn.asp?v=2&n=13].

Sheppard, S. & Jeninson, R. (1997). "Freshman engineering design experiences and organizational framework." *International Journal of Engineering Education, 13* (3), 190-197.

Stiggins, R. J. (1999). "Evaluating classroom assessment training in teacher education programs." *Educational Measurement: Issues and Practice, 18* (1), 23-27.

Yancey, K.B. (1999). "Looking back as we look forward: Historicizing writing assessment." *College Composition and Communication, 50*, 483-503.

Classroom Questions [1]

In 1912, Stevens stated that approximately eighty percent of a teacher's school day was spent asking questions to students. More contemporary research on teacher questioning behaviors and patterns indicate that this has not changed. Teachers today ask between 300-400 questions each day (Leven and Long, 1981).

Teachers ask questions for several reasons (from Morgan and Saxton, 1991):

- the act of asking questions helps teachers keep students actively involved in lessons;
- while answering questions, students have the opportunity to openly express their ideas and thoughts;
- questioning students enables other students to hear different explanations of the material by their peers;
- asking questions helps teachers to pace their lessons and moderate student behavior; and
- questioning students helps teachers to evaluate student learning and revise their lessons as necessary.

As one may deduce, questioning is one of the most popular modes of teaching. For thousands of years, teachers have known that it is possible to transfer factual knowledge and conceptual understanding through the process of asking questions. Unfortunately, although the act of asking questions has the potential to greatly facilitate the learning process, it also has the capacity to turn a child off to learning if done incorrectly. The purpose of this chapter is to provide teachers with information on what types of question and questioning behaviors can facilitate the learning process as well as what types of questions are ineffective.

WHAT IS A GOOD QUESTION?

In order to teach well, it is widely believed that one must be able to question well. Asking good questions fosters interaction between the teacher and his/her students. Rosenshine (1971) found that large amounts of student-teacher interaction promotes student achievement. Thus, one can surmise that good questions fosters student understanding. However, it is important to know that not all questions achieve this.

Teachers spend most of their time asking low-level cognitive questions (Wilen, 1991). These questions concentrate on factual information that can be memorized (ex. What year did the Civil War begin? or Who wrote Great Expectations?). It is widely believed that this type of question can limit students by not helping them to acquire a deep, elaborate understanding of the subject matter.

High-level-cognitive questions can be defined as questions that requires students to use higher order thinking or reasoning skills. By using these skills, students do not

---

[1] Written by Amy C. Brualdi

remember only factual knowledge. Instead, they use their knowledge to problem solve, to analyze, and to evaluate. It is popularly believed that this type of question reveals the most about whether or not a student has truly grasped a concept. This is because a student needs to have a deep understanding of the topic in order to answer this type of question. Teachers do not use high-level-cognitive questions with the same amount of frequency as they do with low-level-cognitive questions. Ellis (1993) claims that many teachers do rely on low-level cognitive questions in order to avoid a slow-paced lesson, keep the attention of the students, and maintain control of the classroom.

Arends (1994) argues that many of the findings concerning the effects of using lower-level-cognitive versus higher-level-cognitive questions has been inconclusive. While some studies and popular belief favor asking high-level-cognitive, other studies reveal the positive effects of asking low-level cognitive questions. Gall (1984), for example, cited that "emphasis on fact questions is more effective for promoting young disadvantaged children's achievement, which primarily involves mastery of basic skills; and emphasis on higher cognitive questions is more effective for students of average and high ability. . ." (p. 41). Nevertheless, other studies do not reveal any difference in achievement between students whose teachers use mostly high level questions and those whose teachers ask mainly low level questions (Arends, 1994; Wilen, 1991). Therefore, although teachers should ask a combination of low-level-cognitive and high-level-cognitive questions, they must determine the needs of their students in order to know which sort of balance between the two types of questions needs to be made in order to foster student understanding and achievement.

HOW TO ASK QUESTIONS THAT FOSTER STUDENT ACHIEVEMENT

In a research review on questioning techniques, Wilen and Clegg (1986) suggest teachers employ the following research supported practices to foster higher student achievement:

- phrase questions clearly;
- ask questions of primarily an academic nature
- allow three to five seconds of wait time after asking a question before requesting a student's response, particularly when high-cognitive level questions are asked;
- encourage students to respond in some way to each question asked;
- balance responses from volunteering and nonvolunteering students;
- elicit a high percentage of correct responses from students and assist with incorrect responses;
- probe students' responses to have them clarify ideas, support a point of view, or extend their thinking;
- acknowledge correct responses from students and use praise specifically and discriminately. (p. 23)

WHAT IS A BAD QUESTION?

When children are hesitant to admit that they do not understand a concept, teachers often try to encourage them to ask questions by assuring them that their

questions will neither be stupid or bad. Teachers frequently say that all questions have some merit and can contribute to the collective understanding of the class. However, the same theory does not apply to teachers. The content of the questions and the manner in which teachers ask them determines whether or not they are effective. Some mistakes that teachers make during the question and answer process include the following: asking vague questions (ex. What did you think of the story that we just read?), asking trick questions, and asking questions that may be too abstract for children of their age (ex. asking a kindergarten class the following question: How can it be 1:00 P.M. in Connecticut but 6:00 P.M. in the United Kingdom at the same moment?)

When questions such as those mentioned are asked, students will usually not know how to respond and may answer the questions incorrectly. Thus, their feelings of failure may cause them to be more hesitant to participate in class (Chuska, 1995), evoke some negative attitudes towards learning, and hinder the creation of a supportive classroom environment.

CONCLUSION

Sanders (1966) stated, "Good questions recognize the wide possibilities of thought and are built around varying forms of thinking. Good questions are directed toward learning and evaluative thinking rather than determining what has been learned in a narrow sense" (p. ix). With this in mind, teachers must be sure that they have a clear purpose for their questions rather than just determining what knowledge is known. This type of question planning results in designing questions that can expand student's knowledge and encourage them to think creatively.

References and Additional Readings

Arends, R. (1994). *Learning to teach.* New York, NY: McGraw-Hill, Inc.

Bloom, B., Englehart, M., Furst, E., & Krathwohl, D. (Eds.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York: David McKay.

Chuska, K. (1995). *Improving classroom questions: A teacher's guide to increasing student motivation, participation, and higher level thinking.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Ellis, K. (1993). Teacher questioning behavior and student learning: What research says to teachers. (Paper presented at the 1993 Convention of the Western States Communication Association, Albuquerque, New Mexico). (ERIC Document Reproduction No. 359 572).

Gall, M. (1970). The use of questions in teaching. *Review of Educational Research*, 40, 707-721.

Gall, M. (1984). Synthesis of research on teachers' questioning. *Educational Leadership*, 42, p. 40-47.

Leven, T. and Long, R. (1981). *Effective instruction.* Washington, DC: Association for Supervision and Curriculum Development.

Morgan, N., and Saxton, J. (1991). *Teaching, Questioning, and Learning.* New York: Routledge.

Rosenshine, B. (1971). *Teaching behaviors and student achievement*. London: National Foundation for Educational Research in England and Wales.

Sanders, N. M. (1966). *Classroom questions: What kinds?* New York: Harper & Row.

Stevens, R. (1912). *The question as a means of efficiency in instruction: A critical study of classroom practice*. New York: Teachers College, Columbia University.

Wilen, W. (1991). Questioning skills for teachers. What research says to the teacher. Third edition. Washington, DC: National Education Association. (ERIC Document Reproduction No. 332 983).

Wilen, W. and Clegg A. (1986). Effective questions and questioning: A research review. *Theory and Research in Social Education*, 14(2), p. 153-61.

Teacher Comments on Report Cards[1]

Several times a year, teachers must complete a report card for each student in order to inform parents about the academic performance and social growth of their child. Schools have a variety of ways to document the progress of students. In a majority of schools, teachers usually assign a number or letter grade to the subject or skill areas. In several schools, mostly elementary schools, teachers write a descriptive narrative of each child's cognitive and social growth. Other schools have teachers indicate whether a student has acquired different skills by completing a checklist.

Despite the fact that schools have different policies concerning the report card's content and format, most teachers are required to include written comments about the student's progress. Considering the amount of students in each classroom, the long span of time needed to complete each report card, and the presence of grade/check marks on the report cards, some may think that comments are nonessential and take up too much of a teacher's time. The purpose of this chapter is to explain why teacher comments on report cards are important, offer suggestions on how to construct effective comments, point out words or phrases to be cautious of using, and indicate sources of information for report card comments.

WHY ARE COMMENTS IMPORTANT?

Grades are designed to define the student's progress and provide information about the skills that he/she has or has not acquired. Nevertheless, grades are often not detailed enough to give parents or the student him/herself a thorough understanding of what the he/she has actually learned or accomplished (Wiggins, 1994; Hall, 1990). For example, if a child receives a B in spelling, a report card comment can inform the parent that the child is generally a good speller; however, she consistently forgets to add an es to plural nouns ending with the letters, s and x. Thus, teacher comments often convey whatever information has not been completely explained by the grade.

Well written comments can give parents and children guidance on how to make improvements specific academic or social areas. For example, the teacher who wrote the previous report card comment on spelling may also wish to include that practicing how to write the different plural nouns at home or playing different spelling games may help the child to enhance her spelling skills.

Words that promote positive view of the student

thorough
caring
shows commitment
improved tremendously
has a good grasp of

The process of writing comments can also be helpful to teachers. Writing comments gives teachers opportunities to be reflective about the academic and social progress of their students. This time of reflection may result in teachers gaining a deeper understanding of each student's strengths and needs.

---

[1] Written by Amy Brualdi

WHAT TYPES OF WORDING SHOULD TEACHERS INCLUDE IN THEIR COMMENTS?

The use of specific comments encourages positive communication between teachers, parents, and students. Written in a positive and informative manner, comments can address a variety of issues while maintaining the while still maintaining the dignity of the child. This is especially important if a child has had difficulty with a particular subject area or controlling his/her behavior over an extended period of time.

Words and Phrases to use to convey that a child needs help

could profit by
requires
finds it difficult at times to
needs reinforcement in
has trouble with

Shafer (1997) compiled a list of "effective" comments from a variety of teachers. The following lists of words and phrases are just a sampling from her publication "Writing Effective Report Card Comments" (p. 42-43).

WORDS AND PHRASES THAT TEACHERS SHOULD BE CAUTIOUS OF USING

When teachers write comments on report cards, they need to be cognizant of the fact that each child has a different rate of social and academic development. Therefore, comments should not portray a child's ability as fixed and permanent (Shafer, 1997). Such comments do not offer any reason to believe that the child will be successful if he/she attempts to improve.

Words to Avoid or Use with Caution

unable
can''t
won''t
always
never

Also, teachers must be sensitive to the fact that their students will read their comments. If negative comments are made, teachers must be aware that those comments may be counterproductive. In addition to the previously mentioned positive comments, Shafer (1997) compiled a list of words and phrases that should be avoided or used with caution (p. 45).

INFORMATION SOURCES TO WHICH TEACHERS SHOULD LOOK WHEN WRITING REPORT CARD COMMENTS

Teachers should have a plethora of sources from which they can derive information on each child to support the comments that are made on each report card. Teachers need these in order to provide specific information on the different strengths and weaknesses of each child. The most commonly used sources of information are examples of student work and test results. In addition to these traditional sources, teachers also use student portfolios as well as formal and informal student observations.

Arter, Spandel, and Culham (1995) define the student portfolio as "a purposeful collection of student work that tells the story of student achievement and growth" (p. 1). A student's portfolio is usually comprised of work that is either the student's best or most exemplary of his/her ability. A portfolio may also contain papers which show the evolution of a particular writing assignment or project. In addition to aiding teachers in keeping track of a student's progress, the portfolio allows the student to chart his/her own academic growth. Because of this, a student should not have many surprises on his report card and will understand how he earned his grades and why different teacher comments were written.

Another rich source of information is the student observation. Student observations often provide important information that is sometimes difficult to derive from the written work of students. These observations allow teachers to make comments on students' daily academic and social behaviors. These should be written about the students' behaviors in a variety of settings: independent work, cooperative learning groups, and playground or nonacademic interaction (Grace, 1992). Grace (1992) suggests that teachers have the following observations for each child: anecdotal records, checklist or inventory, rating scales, questions and requests, and results from screening tests.

References and Additional Readings

Arter J. A., Spandel, V., Culham, R. (1995). Portfolios for assessment and instruction. (ERIC Document Reproduction Service ED388890).

Farr, R. (1991). Portfolios: Assessment in language arts. ERIC digest. (ERIC Document Reproduction Service ED334603).

Grace, C. (1992). The portfolio and its use: Developmentally appropriate assessment of young children. ERIC digest. (ERIC Document Reproduction Service ED351150).

Guskey, T.R. (1996). Reporting on student learning: Lessons from the past-Prescriptions for the future. In Guskey, T.R. (Ed) *Association of Supervision and Curriculum Development Yearbook 1996. Communicating Student Progress*. Arlington, VA: ASCD, pp. 13-24.

Hall, K. (1990). Determining the success of narrative report cards. (ERIC Document Reproduction Service No. 334 013).

Lake, K. and Kafka, K. (1996). Reporting methods in grades K-8. In Guskey, T.R. (Ed) *Association of Supervision and Curriculum Development Yearbook 1996. Communicating Student Progress*. Arlington, VA: ASCD. pp. 90-118

Peckron, K.B. (1996). Beyond the A: Communicating the learning progress of gifted students. In Guskey, T.R. (Ed) *Association of Supervision and Curriculum Development Yearbook 1996. Communicating Student Progress. Arlington*, VA: ASCD pp. 58-64.

Shafer, S. (1997). *Writing Effective Report Card Comments*. New York, NY: Scholastic. [Amazon]

Wiggins, G. (1994). Toward better report cards. *Educational Leadership*. 52(2). pp. 28-37

Essential Skills for Students

Improving the Quality of Student Notes [1]

Much of classroom learning at the secondary and postsecondary levels depends on understanding and retaining information from lectures. In most cases, students are expected to take notes and to review them in preparation for testing of lecture material. Such note-taking may serve a two-fold purpose: as a means of encoding the incoming information in a way that is meaningful for the listener, which serves to make the material more memorable from the outset (encoding function); and as a means of simply storing the information until the time of review (external storage function). Although these two purposes often have been treated as though they were mutually exclusive, several studies (e.g., Maqsud, 1980; Knight & McKelvie, 1986) point to a more complex relationship in which the two vary in their relative importance as a function of the individual, the material, and the review and testing conditions.

DO STUDENTS NEED HELP WITH THEIR NOTES?

Based on several recent investigations, the answer to this question is a resounding "Yes." Of course, some students need more help than others do. Successful students' notes consistently include more of the important propositions, and more propositions overall (though not necessarily more words), than do less successful students' notes (Einstein, Morris, & Smith, 1985). But Kiewra's (1985) summary of the research in this area shows that even successful students generally fail to note many of the important ideas communicated by the lecturer. The best note-takers in these studies (third- year education majors in one study and "A" students in another) included fewer than three quarters of the critical ideas in their notes. First year students fared far worse: their notes contained only 11% of critical lecture ideas.

HOW CAN INSTRUCTORS HELP?

Given that some of the most important information from lectures never is incorporated into students' notes, some means of helping students prioritize their note-taking certainly is in order. A continuum of approaches exists, from providing full or partial lecture notes to modifying one's lecturing style to facilitate students' own note-taking. None of these is optimal in every case. The type of learning (factual versus analytic or synthetic), the density of the information that must be covered, and the instructor's teaching style all should be considered carefully. The merits and drawbacks of each approach are discussed below.

PROVIDING FULL NOTES

Kiewra (1985) reported that students who only review detailed notes provided by the instructor after the lecture generally do better on subsequent fact-based tests of the lecture than do students who only review their own notes. In fact, students who did not even attend the lecture but reviewed the instructor's notes scored higher on such tests than did students who attended the lecture and took and reviewed their own notes.

---

[1] Written by *Bonnie Potts*, American Institutes for Research

Duhaney, D. (Ed.). (2008). *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.

This should not be surprising, because unlike the students' notes, the instructor's notes contain all the critical ideas of the lecture.

One might be tempted, however grudgingly, to conclude that providing students with full transcripts of lectures is the best way to optimize their learning of the material. After all, if the goal is to ensure that they don't miss the important ideas, what better way than to hand each student a full text of the lecture? But Kiewra cites evidence that students remember a greater proportion of the information in their own notes than in provided notes, and that students who take the same amount of time to review both their own and the instructor's notes perform best of all on fact-based tests. Interestingly, the pattern of superior performance with provided notes changes when the test involves higher-order learning (e.g., analysis and synthesis of ideas). In such cases, having the instructor's notes does not produce superior performance.

These results suggest that there is some value in having students participate in the note-taking process, however incomplete their notes may be. A more practical disadvantage to providing full notes is that they may defeat the purpose of the lecture itself. Even if this is not the case (e.g., if lectures serve as opportunities for discussion or other interactive forms of learning), the availability of full notes may encourage absenteeism among students who fail to recognize the additional benefits of attending lectures. These arguments, together with many instructors' understandable objections to preparing and providing full notes, make a compelling case for alternative approaches.

PROVIDING PARTIAL NOTES: THE HAPPY MEDIUM

Several independent investigations (see Russell, Caris, Harris, & Hendricson, 1983; Kiewra, 1985; and Kiewra, DuBois, Christian, & McShane, 1988) have shown that students are able to achieve the most on tests when they are provided with only partial notes to review. Specifically, partial notes led to better retention than did comprehensive (full) notes or no notes, despite the fact that in Russell's study, students expressed an understandable preference for receiving full notes.

Several formats for partial notes have been examined, from outlines, to matrices, to skeletal guides. Of these, the skeletal format has gained the widest support (Hartley, 1978; Russell et al., 1983; Kiewra, 1985). In this format, the main ideas of the lecture are provided, usually including the hierarchical relationships between them (e.g., by arranging them in outline or schematic form), and spaces are left for students to fill in pertinent information, such as definitions, elaborations, or other explicative material, as they listen to the lecture. In Russell's study, students performed especially well with skeletal notes when the test emphasized practical, rather than factual, knowledge of the lecture material. They also remained more attentive during the lecture than did those with other kinds of notes, as evidenced by their higher scores on test-related items presented during each of the four quarters of the lecture period.

Hartley (1978) offered three conclusions from naturalistic research with skeletal notes:

1. Students who get skeletal kinds of notes take about half as many notes of their own, compared to students who are not given notes; yet, students who are given skeletal notes recall more.

2. The amount of space left for note-taking is a strong influence on the amount of notes that students take (i.e., the more space provided, the more notes taken).

3. Although skeletal notes lead to better recall than either the student's own notes or the instructor's notes, the best recall occurred when students received skeletal notes before the lecture and the instructor's detailed notes afterward. (Note the similarity between this finding and that in Kiewra's 1985 study.)

Given the opportunities for analysis and synthesis when one has access to both sets of notes in this way, this result is to be expected.

Ideally, then, instructors would be advised to provide both skeletal notes before the lecture and detailed notes afterward in order to afford their students the maximum benefits. But the disadvantages associated with detailed notes have been discussed above, and given these, it seems unlikely that many educators would choose this option. Certainly, there are also those who would disagree in principle with provision of notes as a remedy for students' difficulties. Instead, it is entirely arguable that emphasis should be placed on helping students improve the quality of their own notes.

HOW CAN STUDENTS' OWN NOTES BE IMPROVED?

Kiewra (1985) offers several suggestions, based on his review of the literature. Some of these call for alterations in the presentation of the lecture. Instructors not only should speak slowly enough to allow students to note important ideas, but also should consider "segmenting" their lectures. Segmenting involves allowing pauses of three to four minutes for every six or seven minutes of lecture. This enables students to devote their attention to listening during the lecture and then to consolidate the important ideas and paraphrase them during the note-taking pauses. During the lecture phase, students need to be given cues not only to the importance of certain ideas, but also to the kinds of elaboration that they are expected to do on these ideas. In certain kinds of classes (e.g., medical school), where the amount of information that must be presented in a given time is relatively great, it may not be possible to segment the lectures, even though students stand to benefit most from segmenting in such cases. A suggested compromise is to keep information density low whenever possible (limiting the presentation of new ideas to 50% of the

Suggestions

- Prepare partial notes for your students.
- Speak slowly so they can write
- Segment your lectures
- Highlight important ideas
- Tell students about expectations
- Encourage students to review their notes
- Encourage students to share notes

lecture time), and to provide skeletal notes in increasing quantity as a function of the lecture's increasing information density.

An additional suggestion by Kiewra (1985) is to encourage students to review not only their own notes, but other sources, such as other students' notes and outside texts. Exposure to a variety of renditions of the same material helps to ensure that the material will be preserved in at least one of the presented forms. It also increases the opportunities for more elaborative processing, as the sources are searched and integrated.

REFERENCES

Einstein, G.O., Morris, J., & Smith, S. (1985). Note-taking, individual differences, and memory for lecture information. *Journal of Educational Psychology*, 77, 522-532.

Hartley, J. (1978). Note-taking: A critical review. *Programmed Learning and Educational Technology*, 15, 207-224.

Kiewra, K.A. (1985). Providing the instructor's notes: An effective addition to student notetaking. *Educational Psychologist*, 20, 33-39.

Kiewra, K.A., DuBois, N.F., Christian, D., & McShane, A. (1988). Providing study notes: Comparison of three types of notes for review. *Journal of Educational Psychology*, 80, 595-597.

Knight, L.J., & McKelvie, S.J. (1986). Effects of attendance, note-taking, and review on memory for a lecture: Encoding versus external storage functions of notes. *Canadian Journal of Behavioral Science*, 18, 52-61.

Maqsud, M. (1980). Effects of personal lecture notes and teacher-notes on recall of university students. *British Journal of Educational Psychology*, 50, 289-294.

Russell, I.J., Caris, T.N., Harris, G.D., & Hendricson, W.D. (1983). Effects of three types of lecture notes on medical student achievement. *Journal of Medical Education*, 58, 627-636.

Helping Children Master the Tricks and Avoid the Traps of Standardized Tests [1]

Adapted with permission from *A Teacher's Guide to Standardized Reading Tests.* Knowledge is Power (1998) by Lucy Calkins, Kate Montgomery, and Donna Santman, Portsmouth, New Hampshire: Heinemann.

Children can improve and change their test-taking habits if they are taught about their misleading work patterns. Teaching children about the traps they tend to fall into may well be the most powerful, specific preparation teachers can give them for the day of the test. By studying the habits of young test takers, we uncovered some of their common mistakes. This chapter lists some of these mistakes and suggests several teaching strategies that may be useful to teachers who are preparing their class to take standardized tests.

USE THE TEXT TO PICK YOUR ANSWER

When it comes to choosing an answer, many children are much more likely to turn to their own memories or experiences than to the hard-to-understand text for their answers. This issue becomes even more difficult when the passage is an excerpt from a text with which the students are familiar. Many new reading tests use passages from well-known children's literature, including those stories that have been made into movies. In this case, many students justify their answers by referring to these movies or their memory of hearing the story when they were younger.

While these personal connections are helpful if the student is at a complete loss for an answer, it's essential for children to understand that relying on opinions, memories, or personal experience is not a reliable strategy for finding answers that a test maker has decided are correct. Clearly, many questions asked on the tests require prior knowledge to answer, but the problem comes when students rely exclusively on that prior knowledge and ignore the information presented in the passage. Some things that teachers may wish to do in order to help their students avoid making this mistake include the following:

- Teach students to underline parts of the passage that might be asked in the questions
- Help children develop scavenger-hunt-type lists of things to look for as they read the passages by having them read the questions first
- Teach students to find out how many questions they can hold in their minds as they read the passage
- Show children how to fill in all the answers on each test booklet page before filling in the corresponding bubbles on the answer sheet
- Teach children ways to mark the passage in order to make it easier to go back to find or check specific parts - these include writing key words in the margins and circling or underlining
- Show students how to use an index card to block out distracting print or to act as a placeholder

[1] Written by Lucy Calkins, Kate Montgomery, and Donna Santman

- Retype familiar or easy text to look as daunting and dense as the test passages to give children confidence and experience in the test format.

SOMETIMES IT'S HELPFUL TO REFER TO YOUR OWN LIFE EXPERIENCES

In the reading comprehension sections of a reading test, children must find evidence in the passages to support their answers. Yet, there are parts of many reading tests where the only things students can rely on are their own previous experiences. In these sections, students are asked to choose the correct spelling of the underlined word or to choose the word whose meaning is closest to that of the underlined word.

Often students prepare for these sections of the tests by taking practice tests and then going over the answers. However, it is highly unlikely that any of the same words would appear on the actual test. Therefore, teachers may wish to impress upon children the importance of creating a context for the variety of words that may be found on the test by relating those words to their own personal reading experiences. In order to facilitate that thinking process, teachers may wish to help children ask themselves such questions as "Have I seen this word before in a book?" "Where have I heard that before?" or "What words or events usually happen around this word?" while they are answering vocabulary or spelling questions.

LEARN TO READ THE QUESTION

It is always assumed that if children have reading troubles, their wrong answers stem from difficulty reading the passages. However, this is not always the case. Sometimes, reading the questions, a much less familiar task, can prove to be the greatest reading challenge for the students. This is because questions such as "How was the central problem resolved?" or "Which statement is NOT true about the narrator?", are not the types of questions children are asking themselves and each other about the books they read.

Studying various types of questions can be a helpful practice to future test takers. This can be done by searching through practice tests and making lists of the types of questions. Although the questions will be different on the day of the test, this exercise may familiarize students with the types of questions that are asked on standardized tests.

CHOOSE THE ANSWER TO THE QUESTION

Sometimes children choose their answer by finding the first answer choice that matches something in the text. Unfortunately, by not considering what the question was actually asking, they are tricked into choosing the wrong answer simply because it may state a fact that was included in the story.

One teaching strategy that can help students avoid this mistake is to present a text with questions in a standardized test format. With a partner, the child should figure out what the different questions are asking, and write down their paraphrased versions. Many times children will be surprised at how different their paraphrasing is from what the question is actually asking. It may be a good practice for teachers to look at the different

paraphrasings with the class and discuss which interpretations would help the members of the class and which would lead them astray. This allows students to strengthen their skills at finding the true meaning of the questions.

RISK AN UNFAMILIAR CHOICE

Frequently, students avoid choosing an answer simply because it contains an unknown word even when they know the other choices are probably wrong. Thus, teachers should advise students not to overlook the possibility that the answer which contains the unfamiliar word may be the correct choice. Teachers often try to teach children a way of narrowing down the answer choices through a process of elimination. Despite the fact that this process can be very helpful, many students eliminate two possibilities and then, from the last two, just sort of pick one. They don't, it seems, try to figure out a reason to choose one over the other. They seem to wrongly assume that the two choices left are equally possible. However, teachers should teach students that thoughtful elimination between the two last possibilities can lead to the correct choice.

CHECK YOUR ANSWERS

After the harrowing ordeal of taking a standardized test, the last thing that students usually want to hear coming from their teacher is "Did you check your answers?" Frequently, the biggest reason kids hate checking answers is because they have only one strategy for doing so: opening their test booklets to the first passage and beginning again. To them, checking answers means taking the test again. However, that does not have to be the case. There are a variety of different strategies that students can use for selectively going back through the test and reconsidering answers. One of these strategies is teaching children to only check the problems of which they were unsure. It is unnecessary to return to questions about which students feel fairly confident. Students can keep track of the troublesome questions while they are actually taking the test. They can do this in several different ways: jotting down the numbers of the questions on a separate sheet of paper, circling the numbers in the test booklet, etc. Students should also know that it is okay to take a short break (stretching in their seats, bathroom / drink break) before going back and checking the answers. This will give them a chance to clear their minds a little bit. Most importantly, students should be taught to attempt to check the answers to the troublesome questions using a new strategy so that they may avoid reusing possibly faulty problem-solving methods.

SETTING THE TONE FOR TEST DAY

Although teachers may do their best to prepare their students for standardized tests, every teacher has stories of children dissolving into tears on the day of tests. Even if their feelings aren't so obvious, all children feel the pressure of doing well. Be sure you don't add to the pressure by over reacting to small deeds of misbehavior or by over emphasizing the fact that today is a testing day.

Suggested Readings

Calkins, L., Montgomery, K. and Santman, D. (1998). *A Teacher's Guide to Standardized Tests. Knowledge Is Power*. Portsmouth, NH: Heinemann.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools*. New York: The Free Press.

Perrone, V. (Ed.). (1991). *Expanding student assessment*. Alexandria, VA: ASCD.

Making the A: How To Study for Tests[1]

Tests are one method of measuring what you have learned in a course. Doing well on tests and earning good grades begin with good study habits. If your goal is to become a successful student, take the time to develop good study habits.

This chapter offers a plan to help you study for tests. It explains how to prepare for and take tests. Techniques for taking essay, multiple choice and other types of exams are reviewed. Although these techniques may help you improve your test scores, other factors, such as class participation, independent projects and term papers also contribute toward grades.

BEFORE THE TEST

Organization, planning and time management are skills essential to becoming a successful student; so start studying as soon as classes begin. Read assignments, listen during lectures and take good classroom notes. Then, reread the assignment, highlighting important information to study. Reviewing regularly allows you to avoid cramming and reduces test anxiety. The biggest benefit is it gives you time to absorb information.

Read difficult assignments twice. Sometimes a second reading will clarify concepts. If you are having difficulty with a subject, get help immediately. Meet with your instructor after class, use an alternate text to supplement required reading or hire a tutor (ask faculty members and other students for referrals).

REVIEW, REVIEW, REVIEW

Plan ahead, scheduling review periods well in advance. Set aside one hour on a Saturday or Sunday to review several subjects. Keep your reviews short and do them often.

- Daily reviews--Conduct short before and after class reviews of lecture notes. Begin reviewing after your first day of class.
- Weekly reviews--Dedicate about 1 hour per subject to review assigned reading and lecture notes.
- Major reviews--Start the week before an exam and study the most difficult subjects when you are the most alert. Study for 2 to 5 hours punctuated by sufficient breaks.

Create review tools, such as flashcards, chapter outlines and summaries. This helps you organize and remember information as well as condense material to a manageable size. Use 3 x 5 cards to review important information. Write ideas, formulas, concepts and facts on cards to carry with you. Study on the bus, in waiting rooms or whenever you have a few extra minutes.

---

[1] Written by Diane Loulou.

Another useful tool is a study checklist. Make a list of everything you need to know for the exam. The list should include a brief description of reading assignments, types of problems to solve, skills to master, major ideas, theories, definitions, and equations. When you begin your final study sessions, cross off items as you review them.

STUDY GROUPS

For some subjects, study groups are an effective tool. Study groups allow students to combine resources; members share an academic goal and provide support and encouragement. Such groups meet regularly to study and learn a specific subject.

To form a study group, look for dedicated students--students who ask and answer questions in class, and who take notes. Suggest to two or three that you meet to talk about group goals, meeting times and other logistics. Effective study groups are limited to five or six people. Test the group first by planning a one-time-only session. If that works, plan another. After several successful sessions, schedule regular meetings.

Set an agenda for each meeting to avoid wasting time. List the material that will be reviewed so members can come prepared. Also, follow a format. For example, begin by comparing notes to make sure you all heard the same thing and recorded important information. Spend 15-20 minutes conducting open-ended discussions on specific topics. Then, test each other by asking questions or take turns explaining concepts. Set aside 5-10 minutes to brainstorm possible test questions.

TAKING AN EXAM

On exam day arrive early and get organized. Pay attention to verbal directions as tests are distributed. Read directions slowly. Scan the entire test, noticing how many points each part is worth and estimate the time needed for individual questions. Before you start answering questions, write down memory aids, formulas, equations, facts and other useful information in the margins.

Check the time and pace yourself. If you get stuck on a question try to remember a related fact. Start from the general and go to the specific. Look for answers in other test questions. Often a term, name, date or other fact you have forgotten will appear somewhere else in the test. Move on to the next question if memory aids do not help. You can always go back to the question if you have time.

TEST-TAKING TIPS FOR DIFFERENT TYPES OF EXAMS

Multiple Choice--Check the directions to see if the questions call for more than one answer. Answer each question in your head before you look at the possible answers. If you can come up with the answer before you look at the choices you eliminate the possibility of being confused by them. Mark questions you can't answer immediately and come back to them later.

When taking a multiple-choice exam guess only if you are not penalized for incorrect answers. Use the following guidelines to make educated guesses.

- If two answers are similar, except for one or two words, choose one of these answers.

- If the answer calls for a sentence completion, eliminate the answers that would not form grammatically correct sentences.

- If answers cover a wide range (5, 76, 87, 109, 500) choose a number in the middle.

For machine-graded multiple-choice tests be certain that the answer you mark corresponds to the question you are answering. Check the test booklet against the answer sheet whenever you start a new section and again at the top of each column.

True-false--If any part of a true-false statement is false, the answer is false. Look for key words, i.e., qualifiers like all, most, sometimes, never or rarely. Questions containing absolute qualifiers such as always or never often are false.

Open book--When studying for this type of test, write down any formulas you will need on a separate sheet. Place tabs on important pages of the book so that you don't have to waste time looking for tables or other critical information. If you plan to use your notes, number them and make a table of contents. Prepare thoroughly for open-book tests. They are often the most difficult.

Short answer/fill-in-the-blank--These tests require students to provide definitions or short descriptions (typically a few words or a sentence or two). Study using flashcards with important terms and phrases. Key words and facts will then be familiar and easy to remember as you answer test questions.

Essay--When answering an essay question, first decide precisely what the question is asking. If a question asks you to compare, do not explain. Standard essay question words are listed next. Look up any unfamiliar words in a dictionary.

Verbs Commonly Used in Essay Questions--Analyze, Compare, Contrast, Criticize, Define, Describe, Discuss, Enumerate, Evaluate, Examine, Explain, Illustrate, Interpret, List, Outline, Prove, State, Summarize.

Before you write your essay, make a quick outline. There are three reasons for doing this. First, your thoughts will be more organized (making it easier for your teacher to read), and you will be less likely to leave out important facts. Second, you will be able to write faster. Third, if you do not have time to finish your answer, you may earn some points with the outline. Don't forget to leave plenty of space between answers. You can use the extra space to add information if there is time.

When you write, get to the point. Start off by including part of the question in your answer. For example, if the question asks, "Discuss the benefits and drawbacks of universal health care coverage to both patients and medical professionals." Your first sentence might read, "Universal health care will benefit patients in the following ways." Expand your answer with supporting ideas and facts. If you have time, review your answers for grammatical errors, clarity and legibility.

Further Reading

Boyd, Ronald T.C. (1988). "*Improving Your Test-Taking Skills*." ERIC Digest No. 101. ERIC Clearinghouse on Tests and Measurement. ED 302 558.

Ellis, David B. (1985). "*Becoming a Master Student*." Fifth Edition. Rapid City, South Dakota: College Survival, Inc.

Mercer County Community College (1992). "Test-Taking Tips." Trenton, N.J. ED 351 597.

Withers, Graeme (1991). *Tackling that test: Everything You Wanted to Know about Taking Tests and Examinations*. Perth: Australian Council for Educational Research.