

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

Copyright © 2002
National Education Association of the United States

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 Assessment 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 representing 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.

Reproduction of any part of this book must include the usual credit line and copyright notice. Address communications to Editor, NEA Student Achievement Department.

Library of Congress Cataloging-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

- c fundamental concepts common to all assessments;
- c essential classroom assessment concepts.
- c 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 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

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

Fundamental Concepts Common to All Assessments

Fundamental Assessment Principles for Teachers and School Administrators²

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.

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

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

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:

- c 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).
- c Student prerequisite knowledge and skills, including test-taking skills.
- c 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. Phey (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. Phey (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.
- 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

- 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.

Essential Concepts for Classroom Assessment

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

Writing Multiple-Choice Test Items¹

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/. 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

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

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.

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.

Stem Checklist

- One point per item
- Doesn't encourage rote response
- Simple Wording
- Short Options

Options Checklist

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

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?

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).

More than factual recall
No superfluous information

STRUCTURE

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

Stem and options related

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 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.

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

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:

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)?

<p><u>OK</u></p> <p>U Different number of option</p> <p>U "None of the above" (sometimes)</p> <p><u>AVOID</u></p> <p>V Typos</p> <p>V Inconsistent grammar</p> <p>V Overlapping distractors</p>

- 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 Classroom¹

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:

- c What concept, skill, or knowledge am I trying to assess?
- c What should my students know?
- c At what level should my students be performing?
- c 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. 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

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:

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

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

- c If possible, have groups of teachers think through the important behaviors included in a task.
- c Express the performance criteria in terms of observable pupil behaviors or product characteristics.
- c Don't use ambiguous words that cloud the meaning of the performance criteria.
- c 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):

- c Checklist Approach When you use this, you only have to indicate whether or not certain elements are present in the performances.
- c 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.
- c 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."
- c 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)
- Rudner, L. and W. Schafer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.

- 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

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

¹ Written by Barbara M. Moskalifer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.

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

- g There appears to be no organization of the document's contents.
- g 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 than 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.

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

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 versus 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 versus Task Specific

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

Use descriptors rather than judgements.

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

the evaluation 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 scoring
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

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

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 than to have many score categories that are difficult or impossible to distinguish.

Each score category should be defined using descriptions of the work rather than 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 subset 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>].

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