DOCUMENT RESUME

ED 373 074 TM 021 811

AUTHOR Wheeler, Patricia H.
TITLE Relative Costs of Various Types of Assessments.
PUB DATE 92
NOTE 11p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Administration; Comparative Analysis; *Constructed Response; *Cost Effectiveness; Cost Estimates; Costs; *Educational Assessment; Estimation (Mathematics); Evaluation Methods; *Multiple Choice Tests; Resource Allocation; Scoring; Selection; *Test Construction; Test Use
IDENTIFIERS *Alternative Assessment; *Performance Based Evaluation

ABSTRACT

Issues of the relative costs of multiple choice tests and alternative types of assessment are explored. Before alternative assessments in large-scale or small-scale programs are used, attention must be given to cost considerations and the resources required to develop and implement the assessment. Major categories of cost to be considered are development and selection costs, administration, and scoring and reporting costs. The relative direct-money costs of multiple-choice, constructed- response, and extended-performance assessments are compared within these three categories. Some examples are given of estimated and actual costs for several major testing programs. It is evident that test users must carefully determine all costs associated with the use of an assessment (direct, indirect, and opportunity) as well as tradeoffs for students and customers. Costs down the road as well as costs of the first year must be considered. The move toward alternative assessment should be done cautiously with careful consideration to short-term and long-term money costs, as well as other costs and benefits associated with this assessment approach. (Contains 20 references.) (SLD)

EREAPA Publication Series No. 92-2

Relative Costs of Various Types of Assessments

Patricia H. Wheeler, M.B.A., Ph.D.

EREAPA Associates
2840 Waverley Way
Livermore, California 94550-1740

1992

The author expresses her appreciation to Geneva D. Haertel of Temple University and EREAPA Associates, and to Celeste Tsuji of Educational Testing Service for their comments and suggestions on earlier versions of this paper, and to Jean Martinson for her editorial changes that improved the readability of the paper.

Relative Costs of Various Types of Assessments

Patricia H. Wheeler, M.B.A., Ph.D.
EREAPA Associates
Livermore, California

While multiple-choice tests have been widely used for the past half-century, alternative assessments continue to gain increased attention for large-scale assessment programs in the United States. Some examples that are being employed in large-scale assessments are essays and writing samples, portfolios, drawings, observations, interviews, work samples, and group projects. Often the term "performance assessment" is used as synonymous with "alternative assessment." However, multiple-choice tests are a means for assessing performance and are considered to be a type of performance assessment. The term "alternative assessment" refers to assessment methods other than multiple-choice. This distinction is not always made in the literature and in the quotations provided in this paper.

The College Board's Advanced Placement Program has been a user of alternative assessments since the mid-1950s, and essays have been part of many testing programs for the past twenty years. Vermont has been using portfolios in its statewide student assessment program. ETS' PRAXIS III will be based on observations of teachers and conferences with these teachers. The California Learning Assessment System (CLAS) will be using group projects as well as new forms of individual assessments for large-scale programs in its new program, scheduled to begin in the spring of 1993. Agencies and states are now beginning to realize that the costs and time associated with such alternative assessments are high, and that the practical issues of funding them need to be confronted (Plato, 1992).

Maeroff (1991) points out that "(s)peed and low cost were the silver bullets that enabled the norm-referenced test--with its multiple-choice responses--to conquer the world of education and hold it in thrall" (p. 275). Alternative assessments have been in use for many years by classroom teachers. They mesh well with instructional activities and provide teachers with much more information about a student's knowledge and skills than do marks gridded into bubbles on answer sheets. However, as Maeroff says, alternative assessments "tend to be a time-consuming, labor-intensive, imprecise exercise in which the expense mounts as nuances are weighed and scoring is done by humans" (p. 275).

Prior to using alternative assessments in large-scale programs as well as smaller ones, attention should have been given to cost considerations and to the resources required to develop and implement them. When asked about problems with implementing authentic assessments, Shepard responded that "cost is a big factor, both for development and scoring" (Kirst, 1991, p. 22). Catterall (1990) points out that we tend to underestimate the true costs of assessments when budgeting for them, to overlook the cost-benefit analysis, and to strive for the political optimum rather than looking at the economic and educational costs and benefits as prime concerns. Not only is there little attention to sound budgeting practices and cost-benefit analysis, but, as pointed out in the next section, few sets of standards for assessments and testing programs even mention cost as a factor to consider.

Standards for Assessments and Programs

Several sets of standards have been issued both to review individual assessments as well as programs. Examples of these are summarized below.

The most widely used standards are those developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1985). These cover technical standards, professional standards for test use, standards for particular applications, and standards for administrative procedures. None of these address cost concerns.

The *ETS Standards for Quality and Fairness* (Educational Testing Service, 1987) covers seven categories of standards, but does not discuss the need to minimize costs while maintaining the quality and fairness standards. The Joint Committee on Testing Practices (1988) has standards for both test developers and test users covering four categories, again none addressing costs: (1) developing/selecting appropriate tests, (2) interpreting scores, (3) striving for fairness, and (4) informing test takers.

Morris, Fitz-Gibbons, and Lindheim (1987) provide a "Test Selector's Screening Questionnaire." Of the 22 features to be rated, none are related to costs. Hambleton and Eignor (1978) developed a proposed set of guidelines for evaluating criterion-referenced tests. They list 39 questions across ten broad categories, again none concerned with costs.

Some criteria and standards do mention costs. For example, Linn, Baker, and Dunbar (1991) identify eight categories of criteria to consider, the last one addressing cost concerns: (1) consequences, (2) fairness, (3) transfer and generalizability, (4) cognitive complexity, (5) content quality, (6) content coverage, (7) meaningfulness, and (8) cost and efficiency. One of the five "APPLE" criteria for assessments of the National Board for Professional Teaching Standards (1991) addresses cost concerns: Administratively feasible, Professionally credible, Publicly acceptable, Legally defensible, and Economically affordable.

Although not developed specifically as standards for assessments, the Joint Committee on Standards for Educational Evaluation (1988) included one under the Feasibility Standards that addresses cost concerns: "Practical Procedures. Personnel evaluation procedures should be planned and conducted so that they produce needed information while minimizing disruption and cost" (p. 71). Such a standard should also apply to individual assessments and assessment programs.

Categories of Costs

Prior to addressing the relative differences in costs between the multiple-choice tests and two major types of alternative assessments, the major categories of costs and the expenses included in these categories should be defined. Three major categories of direct money costs are associated with an assessment program: (1) development/selection, (2) administration, and (3) scoring/reporting.

Development costs for many multiple-choice tests and alternative assessments used in large-scale testing programs are incurred by the test publishers, and amortized over several years through the sale of test materials to users and by fees charged to individual candidates. When such assessments are developed by state or local agencies and employers, funding for such costs should be available prior to the start of the assessment or program development. They can often increase quickly and be higher than anticipated. Test publishers have the option of increasing their market (if the additional costs of doing so are not too high) or using the assessment items and tasks in other instruments and programs. Local and state agencies and employers often do not have such options available to them for covering development costs.

The next major category of direct cost is administration. This occurs once the assessment is developed or selected. For large-scale programs, it can include registering candidates and setting up testing centers. For most programs, it includes obtaining the materials (purchasing or printing), administering the assessments, and accounting for the test materials throughout the period, from receipt of materials through submittal to the scoring agency.

Scoring and reporting cover all activities following the administration of the assessment related to processing of the examinees' response documents (e.g., answer sheets, logs, reports, essays, products). These costs also include providing the results of the assessments to the designated parties (e.g., candidates tested, teachers, licensing boards, employers, college admissions officers, scholarship agencies). Shipment of response documents (e.g., drawing, essays, portfolios) is sometimes done along with score reports and rosters, increasing the cost of reporting.

In addition to these direct money costs, indirect money costs are also incurred within agencies. These can be overhead costs for purchasing and accounts payable staffs, facilities for storage of test materials, rooms for administration of the assessments, and utilities (heat, electricity, water).

Although often overlooked, there are opportunity costs to consider. These costs vary by such factors as the purpose of the assessment, the setting, and the excess resources available for use. The administration of an assessment typically cuts into instructional time or, for employment assessments, into productive work time. The people involved in the administration and scoring of the assessments might better use their time in ways regarded as more valuable to the agency or for their students and customers. The Congress of the United States' Office of Technology Assessment (1992b) says:

> To estimate the opportunity costs, then, requires information or assumptions about the degree to which any particular test is intended as an instructional tool, and information or assumptions about the extent to which the individual teachers use testing as part of their instructional program. (p. 25)

The Office of Technology Assessment (1992b) estimated the 1990-91 cost per student of administering a commercially-published standardized test at about $6.00. However, if teacher costs for administering tests are added, the per student cost increases to $13.00. These figures do not include preparation time.

In reviewing assessments programs in Maryland and Vermont, O'Neil (1992) mentions revamping instruction for new forms of assessment, student preparation time, and administration time as three demands that need to be considered. All three affect opportunity costs. Indirect costs and opportunity costs vary by the type of programs and the test user and, while not covered by this paper, should be considered in the development and use of an assessment or program and in the budgeting process.

Comparison of Relative Costs

Table 1 compares the relative direct money costs of three major types of assessment by areas within three major cost areas. For specific programs, not all of the costs may apply. For example, if one is selecting a published assessment, then costs for conducting pilot testing or assembling assessments may not apply. However, costs for such activities as reviewing test specifications, items and performance tasks, and technical characteristics do apply.

Multiple-choice assessments require an unconstructed response to an item or question, and can be group administered. Alternative assessments fall into two major categories in this table: constructed response and extended performance. Constructed response assessments include essays, drawings, sentence completion, labeling of diagrams, and the working out of mathematical problems. They usually can be group-administered and require little in the way of resources beyond paper and pencil. Extended performance assessments require more resources (e.g., science laboratory equipment, a classroom and group of students to teach, materials and tools for building a prototype, videotaping equipment). In addition, they require more time for the candidate to prepare the assessment (e.g., to develop materials for a portfolio) or for the assessor to administer the assessment (e.g., watching a student perform a series of physical exercises).

The relative comparisons shown in this table are not based on specific data from assessments and testing programs. Rather, they are based on several years of experience with such assessments and testing programs by the author, plus input from colleagues, articles on assessments, test publisher catalogs, and assessment program bulletins and materials. Over the next few years, as specific cost data become available for comparable programs and assessments, more precise comparisons can be made. However, this table may enable test developers and users to anticipate the costs associated with various options being considered more fully.

Costs for many multiple-choice assessments are relatively lower because of the availability of scanners, software, and statistical procedures that have been developed over many years. Also, relative costs can differ markedly, depending on the nature of the specific assessment, and the availability of the expertise and resources needed. The comparisons provided in Table 1 should be considered as rough guidelines for consideration in decision making, not as absolutes.

Local factors should be considered. For example, examination costs may be largely covered by teachers' salaries (Madaus and Kellaghan, 1991). Quellmalz (1984) suggests that costs for training scorers be covered by staff development funds. Some extended performance assessments may require the purchasing or rental of expensive equipment, or the hiring of additional personnel or substitute teachers. Such factors can vary by user, even for the same assessment.

Table 1. Comparison of Relative Direct Money Costs Incurred in the Development/Selection, Administration, and Scoring/Reporting of Multiple-Choice, Constructed Response, and Extended Performance Assessments

Direct Money Cost Areas	Types of Assessment		
	Multiple-Choice	Constructed Response	Extended Performance
Development/Selection:			
Prepare, review, revise test specifications	=	=	=
Write, review, revise items, perf. tasks	+	-	+
Conduct pilot testing	-	-	+
Review items, perf. tasks for technical characteristics, content, format, feasibility, sensitivity	+	-	+
Assemble assessments	+	-	+
Design scoring system	-	+	++
Conduct, analyze data from field testing	-	+	++
Develop, produce final assessment materials	+	-	?
Administration			
Identify, register candidates	=	=	=
Obtain testing facility	-	-	+
Select, train test administrators	-	-	+
Publicize, communicate test information	=	=	=
Purchase test materials	-	-	+
Distribute test materials	-	-	+
Administer assessments	-	-	+
Monitor administrations	-	+	+
Collect, account for materials	-	-	+
Scoring/Reporting			
Code, clean-up, batch answer documents	-	-	+
Select, train scorers	-	+	++
Scan, grade, rate answer documents	-	+	++
Edit documents, quality control, technical checks	-	+	+
Produce score reports	=	=	=
Distribute reports, provide feedback	-	+	+

In this table the symbol = is used when money costs are typically about the same for the three major types of assessments, + or - when the money costs are typically higher or lower compared to the other major types of assessments, and ++ when the money costs are usually much higher compared to the other types of assessments. The question mark indicates that there can be large variations, depending on the specific assessment.

Examples of Estimated and Actual Costs

Regarding costs, O'Neil (1992) comments that "(a)lthough estimates vary on the costs of traditional machine-scored, multiple-choice tests versus performance assessments, some experts say performance assessments are likely to be at least two or three times more expensive per student" (pp. 17-18).

The Congress of the United States' Office of Technology Assessment (1992a) provides the following summary of estimated costs for alternative assessments versus multiple-choice tests:

> The costs of performance assessment represent a substantial barrier to expanded use. Performance assessment is a labor-intensive and therefore costly alternative unless it is integrated in the instructional process. Essays and other performance tasks may cost less to develop than do multiple choice items, but are very costly to score. One estimate puts scoring a writing assessment as 5 to 10 times more expensive as scoring a multiple-choice examination, while another estimate, based on a review of several testing programs administered by ETS, suggests that the cost of assessment via one 20- to 40-minute essay is between 3 to 5 times higher than assessment by means of a test of 150 to 200 machine-scored, multiple-choice items. Among the factors that influence scoring costs are the length of time students are given to complete the essay, the number of readers scoring each essay, qualifications and location of readers (which affects how much they are paid, and travel and lodging costs for the scoring process), and the amount of pretesting conducted on each prompt or question. The higher these factors, the higher the ratio of essay to multiple-choice costs. The volume of essays read at each scoring session has a reverse impact on cost--the greater the volume, the lower the per item cost. (p. 243)

Based on 1992-93 bulletins and registration forms for several major testing programs, there are wide variations in candidate fees, reflected in part by the type of examination. The Graduate Management Admission Test (GMAT) is entirely multiple choice and costs $42 per candidate. Both the Pre-Professional Skills Tests (PPST, all three parts) and the Law School Admission Test (LSAT) have multiple-choice and essay sections; the candidate fees are $65 and $71 respectively. All three programs require a half day of testing.

The Graduate Record Examination (GRE) General Test consists of three multiple-choice sections, and is available in paper-and-pencil and computer-based versions. The first version costs $45 per candidate, whereas the candidate fee for the computer-based version is $90. In part, this difference reflects the administrative conditions (large group versus small group or one-on-one in a facility with computers).

The College Board's Test of English as a Foreign Language/Test of Written English (TOEFL/TWE) contains both multiple-choice sections and an essay; the fee per candidate in 1992-93 is $35. However, the Test of Spoken English (TSE), which requires audio taping of the candidate's responses, costs $80 to $110 per candidate. This illustrates the higher costs associated with extended performance assessments, as compared to multiple-choice and essay assessments.

A comparison of scoring service costs in the 1992 catalogs of two major test publishers illustrates the variations in the costs for processing various types of answer documents. For example, for the full battery of the Comprehensive Tests of Basic Skills (CTBS/4), CTB Macmillan/McGraw-Hill charges $1.37 for a machine scorable answer sheet for levels 14-22 and $2.53 for a test booklet for levels 12-13. Their charges for the scoring of essays are: $4.10 for holistic only, $4.10 for analytic only (one point), $5.20 for holistic and one-point analytic, and $1.16 for each additional analytic scoring point.

The Psychological Corporation charges about 20% more to process a machine-scannable test booklet as compared to an answer sheet (e.g., $2.74 versus $3.37 for the same test). Their charge of $4.90 is for holistic scoring of an essay and $4.30 is added for analytic scoring.

It is clear from the examples above that variations in the assessment methods and scoring affect costs and need to be considered in the planning of any assessment program.

Summary

Based on their longer term experience with performance assessment programs in England, Nuttall (1992) suggests two lessons: "first, the cost of performance assessment, both financially and in terms of the time of teachers, is immense; and, second, despite all the care and effort, some will still not view it as rigorous enough" (p. 57).

Herman (1992) warns us that "(w)ith more labor-intensive, performance-based assessments, greater attention will need to be given to efficient data collection designs and scoring procedures" (p. 76).

Maeroff (1991) reminds us that "(w)hile it may be possible to be systematic about alternative assessments, there are ultimately no quick and easy ways to rate large numbers of performance-based tasks or portfolios or interviews or exhibits or even essays" (p. 275).

In their report, *Raising Standards For American Education*, the National Council on Education Standards and Testing (1992) discusses costs as an argument against a national system of assessment in the United States. "The costs of this system in terms of teacher development, task development, administration, scoring, and validation are so high as to be insupportable" (p. F-12).

Test users must carefully determine all costs associated with the use of an assessment--direct, indirect, and opportunity--and the tradeoffs for their students and customers. Not only must the costs for this first year be considered, but we must look at costs down the road several years. Again, many of the direct costs associated with multiple-choice testing are relatively lower because of the development of scanners, software, and statistical methodologies. However, these costs were not low originally. Computer-based methods for grading some constructed-response assessments on a large-scale basis are very close to implementation. Reliance on people to do such grading over many years can have its drawbacks; teachers don't want to spend every summer grading essays, and, after many years, it can be a challenge to locate enough essay readers and assessment graders.

The move toward alternative assessments should be done cautiously and with careful consideration to short-term and long-term money costs, as well as other costs and benefits associated with such an approach to assessment.

References

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

Catterall, James S. (1990, November). *Estimating the costs and benefits of large-scale assessments: Lessons from recent research* (CSE Technical Report 319). Los Angeles, CA: University of California at Los Angeles, Center for Research on Evaluation, Standards, and Student Testing.

Congress of the United States, Office of Technology Assessment. (1992a). *Testing in American schools: Asking the right questions* (Report No. OTA-SET-519). Washington, DC: Author.

Congress of the United States, Office of Technology Assessment. (1992b, February). *Summary: Testing in American schools: Asking the right questions* (Report No. OTA-SET-520). Washington, DC: Author.

Educational Testing Service. (1987). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Hambleton, Ronald K.; & Eignor, Daniel R. (1978, Winter). Guidelines for evaluating criterion-referenced tests and test manuals. *Journal of Educational Measurement, 15*(4), 321-327.

Herman, Joan L. (1992, May). Synthesis of research: What research tells us about good assessment. *Educational Leadership, 49*(8), 74-78.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators.* Newbury Park, CA: Sage Publications, Inc.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: American Psychological Association.

Kirst, Michael. (1991, March). Interview on assessment issues with Lorrie Shepard. *Educational Researcher, 20*(2), 21-23, 27.

Linn, Robert L.; Baker, Eva L.; & Dunbar, Stephen B. (1991, November). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Madaus, George F.; & Kellaghan, Thomas. (1991). Student examination systems in the European community: Lessons for the United States. In Gerald Kulm & Shirley M. Malcom (Eds.), *Science assessment in the service of reform* (pp. 189-232). Washington, DC: American Association for the Advancement of Science.

Maeroff, Gene I. (1991, December). Assessing alternative assessment. *Phi Delta Kappan, 73*(4), 272-281.

Morris, Lynn Lyons; Fitz-Gibbons, Carol Taylor; & Lindheim, Elaine. (1987). *How to measure performance and use tests*. Newbury Park, CA: Sage Publications, Inc.

National Board for Professional Teaching Standards. (1991). *Toward high and rigorous standards for the teaching profession: Initial policies and perspectives of the National Board for Professional Teaching Standards* (3rd ed.). Detroit, MI: Author.

National Council on Education Standards and Testing. (1992, January 24). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people*. Washington, DC: U.S. Government Printing Office.

Nuttall, Desmond L. (1992, May). Performance assessment: The message from England. *Educational Leadership*, *49*(8), 54-57.

O'Neil, John. (1992, May). Putting performance assessment to the test. *Educational Leadership*, *49*(8), 14-19.

Plato, Kathleen. (1992, September). The politics of assessment reform: Implications for educators. *NASSP Bulletin*, *76*(545), 41-49.

Quellmalz, Edys S. (1984, Spring). Designing writing assessments: Balancing fairness, utility, and cost. *Educational Evaluation and Policy Analysis*, *6*(1), 63-72.