C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Educational Assessment: Expanded Expectations and Challenges

CSE Technical Report 351

Robert L. Linn
CRESST/University of Colorado at Boulder

January 1993

Invited Address: Thorndike Award from Division 15, Educational Psychology, of the American Psychological Association presented at the annual meeting of the American Psychological Association, Washington, DC, August 16, 1992.

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported in part under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

Educational Assessment: Expanded Expectations and Challenges

Abstract

Current national efforts to expand the role of educational assessment and radically change the nature of the assessment are discussed. Rationales for expectations that a new national examination system would serve as a lever of educational reform are analyzed. Challenges posed for educational measurement by the proposed heavy reliance on complex, performance-based assessment are examined in terms of the needed comprehensive validation research. Particular attention is given to existing generalizability evidence and the implications of high degree of task specificity in performance for the design of an assessment system.

Educational Assessment: Expanded Expectations and Challenges

Robert L. Linn

CRESST/University of Colorado at Boulder

The recent report "Testing in American Schools: Asking the Right Questions" (U.S. Congress, Office of Technology Assessment, 1992), prepared at the request of Congress by the Office of Technology Assessment (OTA), documents the central role of educational testing in recent national debates about educational reform. Although the current emphasis on assessment has some important new features that I will discuss in some detail, it is hardly a novelty for testing and assessment to figure prominently in policy makers' efforts to reform education. As Madaus (1985) observed several years ago, "testing is the darling of policy makers across the country" (p. 5). Similar statements could have been made at various times during the past century and a half, most notably during periods when the schools were under attack and reformers sought to demonstrate the need for change.

As has been true of previous reform efforts, assessment is central to the current educational reform debate for at least two reasons. First, assessment results are relied upon to document the need for change. Second, assessments are seen as critical agents of reform. Indeed, Petrie (1987) went further when he argued that "It would not be too much of an exaggeration to say that evaluation and testing have become *the* engine for implementing educational policy" (p. 175, emphasis in the original).

The primary focus of this paper is that educational policy makers are keenly interested in educational assessment and that their greatly expanded, and sometimes unrealistic, expectations, together with the current press for radical changes in the nature of assessments, represent major challenges for educational measurement. First, however, I will discuss the attraction that assessment results have for policy makers.

Barometer of Educational Quality

Educational assessments are often rather naively expected to serve as a kind of impartial barometer of educational quality. Such an expectation makes assessment results of particular interest and value to various types of policy makers. In this regard, policy makers have often pointed to educational assessment results in order to demonstrate educational shortcomings.

The OTA report provides a brief recounting of this history of testing in American schools from the time that Horace Mann introduced written examinations in the mid-19th century. The OTA report (U.S. Congress, Office of Technology Assessment, 1992) summarized the view that tests could document the need for change as follows:

> The idea underlying the implementation of written examinations, that they could provide information about student learning, was born in the minds of individuals already convinced that education was substandard. This sequence—perception of failure followed by the collection of data designed to document failure (or success)—offers early evidence of what has become a tradition of school reform and a truism of student testing: tests are often administered not just to discover how well schools or kids are doing, but rather to obtain external confirmation—validation—of the hypothesis that they are not doing well at all. (p. 108, emphasis in the original)

More recent examples where test results have been used to document the need for reform are plentiful, but two will suffice to make the point. The June 6, 1991 press conference for the release of the first state-by-state results of the National Assessment of Educational Progress (NAEP) in mathematics was used by Secretary of Education Lamar Alexander to say that the results should serve as a "wake up America call." According to Secretary Alexander, "The big news is: None of the states is cutting it." In the second example, President Bush relied on a much less relevant source of test data for purposes of drawing inferences about schools when he made the following comments regarding the release of SAT results in the fall of 1991. "Last week we learned SAT scores had fallen again. Scores on the Verbal SAT have tumbled to the lowest level ever. These numbers tell us our schools are in trouble" (cited by Jaeger, 1992).

The remainder of this paper could be used to analyze what is wrong with this type of use of SAT results, but that is not my focus today and, in any event,

that has already been done on several occasions (e.g., Linn, 1987; Wainer, 1986). It is worth noting, however, that despite the negative picture, there is considerable evidence regarding improvements in achievement that the critics choose not to acknowledge. For example, even if it were reasonable to use aggregate SAT results as an indicator of educational progress, results presented by Carson, Huelskamp, and Woodall (1991) and discussed by Berliner (1992), and by Jaeger (1992), demonstrate that the global comparisons of cross-sectional results for different years hide more than they reveal. Jaeger's (1992) discussion is particularly telling:

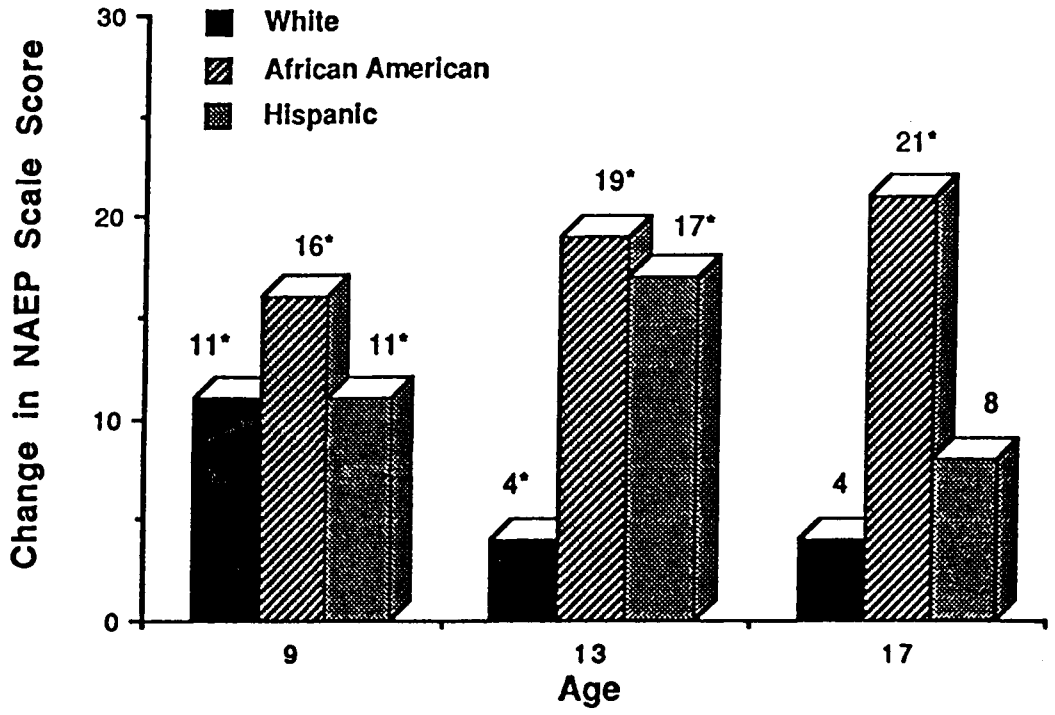
> As Carson and his colleagues (Carson, Huelskamp and Woodall, 1991) have illustrated, the SAT score decline is a perfect example of Simpson's Paradox. Although the mean score of the total SAT test-taking population has declined, the mean score of every major racial and ethnic group that composes the population of SAT test-takers has increased during the last 15 years. Comparison of the mean performances of the population of SAT test-takers today and the population of test-takers fifteen or twenty years ago, unadjusted for changes in the test-taking population, can only produce seriously misguided conclusions. (pp. 7-8)

Turning to the potentially more informative NAEP results, it should be noted that Secretary Alexander's analysis of the cross-sectional results in mathematics does not take into account trend data that show improvements in mathematics achievement especially during the 1980s (see Jaeger, 1992; Linn, in press; Mullis, Dossey, Foertsch, Jones, & Gentile, 1991). As can be seen in Figure 1, the performance of African-American students on the NAEP mathematics scale was significantly higher at all three age levels in 1990 than it was in 1978. The performance of White and Hispanic students also was higher in 1990 than in 1978 at all three age levels, but the differences were not significant for 17-year-olds. For the 1990 bridge samples, the within-grade standard deviations ranged from 31 to 33. In other words, a difference of about 16 points represents an effect size of about .5. Thus, most of the differences shown in Figure 1 are large enough to be of practical as well as statistical significance.

Although policy makers seeking reform make great use of test results to argue the case that education is inadequate, others with an interest in the status quo, of course, emphasize quite different results. As was demonstrated

Figure 1

Changes in Average Performance on the Overall NAEP Mathematics Scale Between 1978 and 1990 (Based on Mullis, Dossey, Foertsch, Jones, & Gentile, 1991)



* Significant Increase in Mean ($p < .05$)

in articles on the Lake Wobegon effect (e.g., Cannell, 1987, 1988; Linn, Graue, & Sanders, 1990), reports comparing student achievement to outdated national norms for the form of a test that has been used year after year in a state or school district have provided misleadingly positive impressions of student achievement in individual states and districts.

Figure 2 shows results from a study that I have worked on in collaboration with Dan Koretz, Lorrie Shepard, Steve Dunbar, Freddy Hiebert, and Bobbie Flexer (e.g., Koretz, Linn, Dunbar, & Shepard, 1991). The operational test results in two districts using different norm-referenced tests in relatively high-stakes testing programs are contrasted with results of administrations of alternative tests that were constructed to cover the same content objectives and then equated to the district norm-referenced tests using samples of students from other districts where the norm-referenced test in question was not used. There is good generalization regarding mean student performance in only one of the four contrasts (mathematics in district A). In the other three contrasts, the operational test provides an inflated impression of student achievement in comparison to the alternative test.

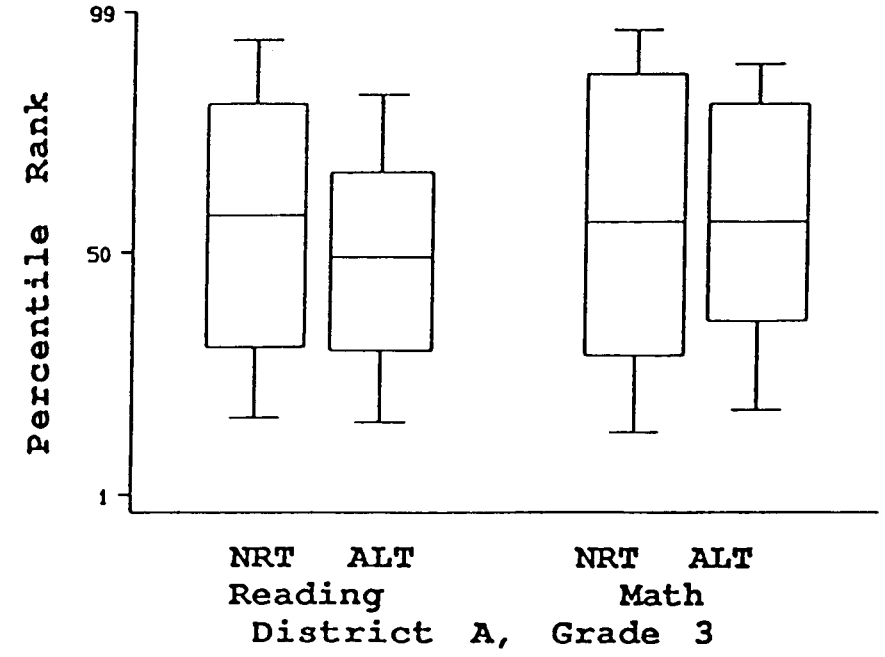
The possibly obvious point is that considerable caution is needed in using achievement test results to draw inferences about the quality of education. Despite the pitfalls, however, policy makers attempting to provide support for current practice, as well as those trying to undermine it, continue to rely heavily on test results to make their case.

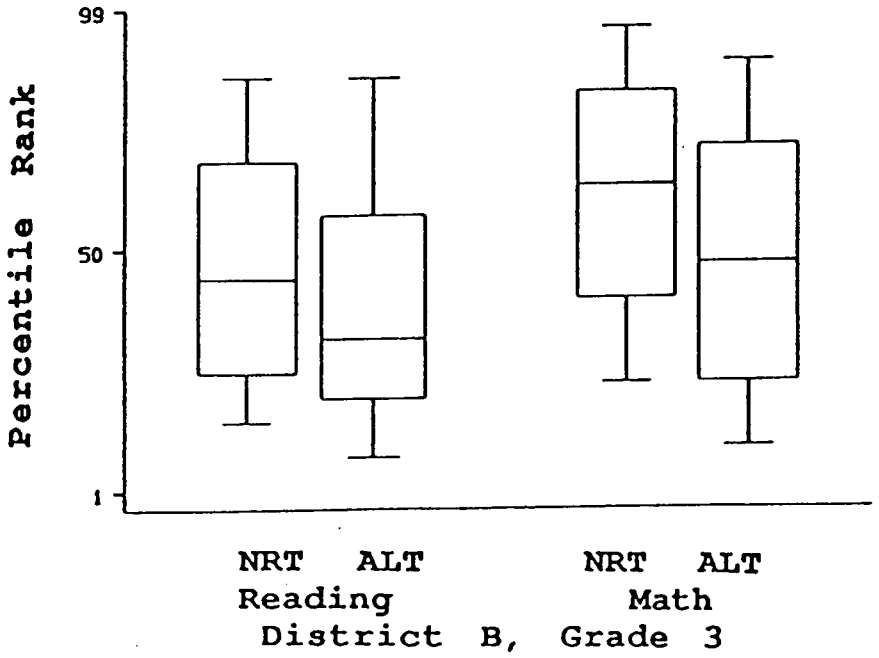
Instrument of Reform

Issues regarding the uses and misuses of educational assessment results to draw inferences about the quality of education are worthy of more detailed discussion, but the focus of the remainder of this paper is on a second use of educational assessment. That is the use of assessments not just as a barometer of educational conditions but as an instrument of reform. Educational assessments are expected not only to serve as a monitor of educational achievement but to be powerful tools of educational reform.

Figure 2

Box-and-Whisker Plots of Percentile Rank Distributions of Operational Norm-Referenced Tests (NRT) and Alternative Tests (ALT) for Two School Districts with High-Stakes Testing Programs





During the 1970s and 1980s state after state turned to testing as a central element in educational reform efforts, first with the introduction of minimum competency testing requirements and then with increased requirements and accountability systems with increased stakes. Although the states are still key players, much of the recent debate about the role of assessment as an instrument of reform is now taking place at the national level. Proposals for national tests or a national system of examinations have flourished during the past couple of years.

Assessment is a core component of (a) the Bush Administration's *America 2000* proposals (U.S. Department of Education, 1991); (b) the National Education Goals Panel's desire to use assessments not only to monitor progress but to help bring it about; and (c) the report of the Secretary of Labor's Commission on Achieving Necessary Skills (SCANS) (U.S. Department of Labor, 1991). Non-governmental groups are also calling for national assessments either in the form of a national test as proposed by Educate America or a national system of examinations keyed to common national standards as proposed by the New Standards Project (see, for example, Resnick, 1992).

In response to the rapid expansion of interest in a national test or system of examinations, Congress created the National Council on Education Standards and Testing (NCEST) in June, 1991 and charged the council to:

- advise on the desirability and feasibility of national standards and tests, and
- recommend long-term policies, structures, and mechanisms for setting voluntary educational standards and planning an appropriate system of tests. (National Council on Education Standards and Testing, 1992, p. 1)

The NCEST report, submitted in January 1992, concluded "that national standards and a system of assessments are desirable and feasible mechanisms for raising expectations, revitalizing instruction, and rejuvenating educational reform efforts for all American schools and students. Thus the National Council on Education Standards and Testing endorses the adoption of high standards and the development of a system of assessments to measure progress toward those standards" (National Council on Education Standards and Testing, 1992, p. 8).

Both the House and Senate have bills pending that deal with aspects of the NCEST report, albeit in different ways. It is unclear, particularly in view of the upcoming election, however, what, if any, legislation will be forthcoming. In the mean time, it seems likely that some of the NCEST recommendations will be pursued, at least on a limited scale, by the National Education Goals Panel. Similar ideas are also being promoted by the New Standards Project and various clusters of states working together with organizational assistance from the Council of Chief State School Officers. Hence, some of the key elements of the NCEST proposal seem worthy of analysis.

The new system of assessments proposed in the NCEST report has two components: "individual student assessments, and large-scale sample assessments such as the National Assessment of Educational Progress" (National Council on Education Standards and Testing, 1992, p. 4). Both assessment components would be closely aligned to the proposed national content standards and would provide the basis for establishing performance standards (i.e., levels of competence or desired levels of student performance on the assessments). The individual student assessments would not be a single test, but would involve multiple methods. Encouragement was given for the use of performance-based assessments. As conceptualized, states or groups of states might devise or adopt different approaches to assessments that would be linked to the national standards in ways that it is hoped would lead to comparable results.

A critical assumption underlying the NCEST recommendation, as well as related proposals in *America 2000*, the SCANS report, and from the New Standards Project, is that the establishment of clearly defined high standards and assessments with associated rewards and sanctions will motivate students and teachers to put forth greater effort. A second assumption is that the negative effects associated with previous reform efforts based on high-stakes uses of standardized tests can be overcome by the introduction of assessments, particularly performance-based assessments, that are closely aligned to national content and performance standards. Both of these assumptions deserve more detailed analysis. They also pose challenges to educational researchers who would contribute to the design and validation of an assessment system along the lines proposed by NCEST and elsewhere.

Motivation. The motivational assumption is evident in all the proposals for either a national test or a national system of assessments. The NCEST report, for example, is clear that enhanced motivation is a central goal. "The new assessments should challenge all students and educators to do their best, open up new opportunities for students, and provide real incentives to improve the quality of America's schools" (National Council on Education Standards and Testing, 1992, p. 28).

Will high standards and assessments motivate students and teachers to work harder? There is certainly no shortage of testing in the schools now. Many students wishing to go to college still get SAT scores and ACT scores below those needed to be admitted to the college of their choice. Such results apparently have not motivated students to achieve at the levels expected. Why should a new testing system be a better motivator?

The answer commonly given to this question has two parts. First, most current, externally-imposed tests are *not* closely linked to the instructional goals of the schools. Second, with a few exceptions, such as Advanced Placement tests, results on current tests that have high standards associated with them are not closely tied to real and visible sanctions and rewards. Of course, minimum competency tests have sanctions associated with them, but the standards are low in comparison to the "world-class" standards characterized in the current rhetoric about national examination systems.

The lack of linkage between the curriculum and externally-imposed tests in this country represents one of the major differences from examinations used in other industrialized countries. Most externally-imposed tests in the U.S. are designed to measure generic skills that are decoupled from any specific curriculum. Despite considerable use of coaching schools, the SAT is not designed as a test that students are expected to study for. Many see this decoupling as a major weakness of our testing system (e.g., Resnick & Resnick, 1992). For example, after noting some of the negative effects of examination systems in other industrially-developed nations, Smith, O'Day, and Cohen (1990) go on to argue that there are also positive lessons to be learned from those systems. "The first and central lesson is this: If exams are used to motivate students to be more serious about their studies, then examinations' content must be closely tied to the curriculum frameworks that are used to teach students" (Smith et al., 1990, p. 41). This linking of

assessment to the curriculum is, in my opinion, one of the most positive aspects of the proposed test-based reforms.

The second aspect, rewards and sanctions, on the other hand seems more problematic. Nonetheless, rewards and sanctions for individual students are seen as a key to student motivation. *America 2000*, for example, suggests several mechanisms for increasing the impact of the proposed American Achievement Tests on student, teacher, and parent efforts. The plan calls for the award of Presidential citations to "students who do well on the American Achievement Tests" (U.S. Department of Education, 1991, p. 14) and for the reward of Presidential Achievement Scholarships. It is also indicated that colleges and employers will be urged to use the test results in making admissions and hiring decisions. For parents, the results are supposed to provide information about how schools are doing so they can make choices among schools for their children.

The NCEST report includes a caution that high stakes should not be attached to the assessment results before the "... qualities of validity, reliability, and fairness have been addressed" but goes on to conclude that "... assessments eventually could be used for such high-stakes purposes as high school graduation, college admission, continuing education, or certification for employment. Assessments could also be used by states and localities as the basis for accountability" (National Council on Education Standards and Testing, 1992, pp. 27-28).

The importance of incentives as motivators for better performance is clearly articulated by Smith et al. (1990):

> First and foremost, we argue that national examinations used for either student or system accountability will be legitimate and useful if they are based on the national curriculum frameworks. To motivate students, the exams would provide incentives to excel, both by offering challenging content that requires effort and attention from students at all ability levels and through real-life rewards for good performance. For college-bound students, these rewards might be related to university admission; for work-bound students, good exam scores might mean better job prospects. (p. 42)

In contrast to their presumed motivational benefits, it is important to recognize that high-stakes tests have many potential down-sides. The negative effects of existing high-stakes tests have been well documented. Some of these, such as the focus on basic skills and the drill and practice on factual knowledge in formats mimicking multiple-choice tests, may be avoided by the introduction of performance-based approaches to assessment that rely more heavily on extended tasks. Other negatives, however, are less easily eliminated.

Students who believe that they have a reasonable chance of meeting the standards, and who believe that there are real rewards for doing so, may indeed be motivated to work harder. On the other hand, as has been shown by research linking minimum competency testing and dropout rates (Catterall, 1987), students who see the standards as beyond their reach or who do not believe that outcomes that they value will be associated with meeting the standards are not likely to be motivated by new requirements. The tendency for students to give up may be exacerbated by school and district actions designed to make the system look good. If the system is made to look bad, low-performing students may be rejected from the system just as they have been in other test-based accountability systems.

Validity. If the NCEST provision for obtaining supporting validity evidence is taken seriously, then validators will have a full agenda, for as Cronbach (1988) has argued, "validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences" (p. 6). Presumed consequences such as enhanced student and teacher effort are central to the argument for a national system of assessments. Perceived negative consequences of existing high-stakes testing programs that rely on multiple-choice tests are critical to the arguments for a system of assessments that relies on performance-based assessment tasks that proponents believe would provide better instructional targets.

Consequences should be just as central to the evaluation of any new assessments. Messick's (1989) discussion of the consequential basis of validity provides a convincing case that consequences should be a major focus of the validation of the uses and interpretations of any measure (see, also Cronbach, 1988; Linn, Baker, & Dunbar, 1991; Messick, 1992). The need to obtain evidence

regarding consequences is especially compelling for performance-based assessments such as those envisioned by NCEST, SCANS and the National Education Goals Panel, however, because particular intended consequences are an explicit part of the rationale for the assessment system.

Plans for an evaluation of consequences should start with the effects that the assessment system is intended to have. The assessments are expected to be used by, and have an impact on, schools, colleges and employers. They are expected to have an impact on what and how teachers teach. And, they are expected to motivate students to put substantially greater effort into their school work. Each of these intended consequences needs to be evaluated.

Does the assessment lead teachers to change the nature of assignments that are given to students? How similar are classroom activities and homework assignments to the tasks that are included in the assessment system? Is the allocation of time to different content domains altered as the result of the assessment and, if so, is the re-allocation deemed to be a desirable one?

If a version of the proposed standards and associated assessments is introduced, it will be critical to track the ways in which colleges and employers interpret and use the assessment results for at least two reasons. First, the uses and interpretations will need to be justified. Second, because college and employer uses are expected to be beneficial not only for the institutions that use them but as motivators of student performance in school, it will be important to know the prevalence and nature of those uses. It also will be valuable to monitor the level of student awareness of, and beliefs about, the uses that colleges and employers make of the assessment results.

Justification of use by employers will depend, in part, on the impact the use has for the employment of protected groups. If the use of the assessment has an adverse impact on the employment of minorities or women, then evidence will need to be obtained to support the claim that the assessed competencies are job related and consistent with business necessity.

Evidence regarding the impact of the assessments on student motivation and behavior arguably will be the most difficult as well as the potentially most important to obtain. Questionnaire surveys and interviews of students regarding their beliefs about the importance of high school records and

performance on assessments could provide useful information. Evidence regarding changes in student behavior (e.g., time spent studying and performance on day-to-day classroom assignments) would provide more compelling evidence.

It will be important that validation research focusing on the effects of the assessment system on student behavior address a wide range of both positive and negative potential effects. It might be shown, for example, that the assessment increased the effort of some, but not all, students. Such an outcome would provide positive support for the value of the assessment, but that positive support would need to be weighed against evidence regarding the possibly negative impact of the assessment on other students. Do some students give up because they believe that performance required for certification is beyond their reach? Are dropout rates increased?

Fairness. Although often discussed as an independent topic, the fairness of an assessment is an essential aspect of an overall judgment regarding validity in the sense articulated by Messick (1989). Fairness clearly is a major consideration in judgments regarding the appropriateness of the uses and interpretations of an assessment.

It would be a serious mistake to assume that performance-based assessments are somehow immune to problems of bias or adverse impact. Because there are large between-group differences in educational opportunity, there are also likely to be differences in results on performance-based assessments, at least in the short run. Indeed, some research suggests that the gap between the performance of underserved minority groups and the majority group may be as large or larger with performance-based measures as with traditional tests (Linn et al., 1991).

Resnick (1990) has argued that the real issue of bias for the type of performance-based assessments that she has championed is differential access to opportunities to learn. Of course, this argument also has been made with regard to standardized tests. The argument has considerable merit in both contexts, and it is an important message to convey, but it is even more important to change the degree of bias in access to opportunity that now exists. Without a fundamental change in educational opportunities for underserved minorities, group differences that are all too familiar on current tests can be

expected to continue on a new set of performance-based assessments. The resulting disparate impact on minority students will not only undermine the system but demonstrate a failure to achieve the goal of providing better education for all students.

An interpretation of fairness in terms of access to instructional opportunities implies the need to evaluate the degree to which students are provided with the needed instructional supports to prepare them for the assessment. A system for monitoring instructional experiences as well as student outcomes may be essential if the assessments come to have major importance in employment and college admissions decisions.

The NCEST report provides support for the idea that student performance standards need to be accompanied by school delivery and system performance standards. The report of the NCEST Standards Task Force argued that school delivery standards should provide the means for "determining whether the school 'delivers' to students the 'opportunity to learn' well material in the *content standards*" (National Council on Education Standards and Testing, 1992, p. E-5). The Task Force report goes on to elaborate this idea of school delivery standards by listing a set of tough questions:

> Are the teachers in the school trained to teach the content of the standards? Does the school have appropriate and high quality instructional materials which reflect the *content standards*? Does the actual curriculum of the school cover the material of the *content standards* in sufficient depth for all students to master it to a high standard of performance? (p. E-5)

Such questions provide a challenging research agenda for the proposed national system of examinations. It should be noted, however, that school delivery and system performance standards are not ideas that have been accepted by many proponents of national examination systems. Indeed, the school delivery standards aspect of the proposed legislation has been opposed by Secretary of Education Lamar Alexander.

If delivery standards are not a part of the design of an examination system with rewards and sanctions, they are likely to become a legal issue. As was demonstrated by the *Debra P. v. Turlington* case[1], assessments may be

[1] Debra P. v. Turlington, 474 F. Supp. 244, 265 (M.D. Fla. 1979); 644 F.2d 397 (5th Cir. 1981).

subject to legal challenge on due process grounds unless there is evidence that students have been given a reasonable opportunity to learn the skills or competencies assessed. In Debra P. the Fifth Circuit panel concluded that there was a due process violation in the Florida's introduction of a minimum competency test requirement for high school graduation. The Court's reasoning in support of this ruling is instructive. "We believe that the state administered a test that was, at least on the record before us, fundamentally unfair in that it *may* have covered matters not taught in the schools of the state."[2]

Demonstration that schools have provided students with an adequate opportunity to prepare for the assessments is obviously a tall order. The importance of taking the challenge of documenting that students are given adequate opportunity goes well beyond the potential legal demands suggested by Debra P., however.

Performance-based assessments. As is implicit in some of the comments about using assessments to motivate students, an important assumption of the proposed systems of examinations is that the deleterious effects associated with previous efforts at test-based accountability can be overcome by switching to performance-based assessments that are closely linked to curriculum frameworks. The California State Department of Education (1990) report on its education summit meeting clearly reflects this view:

> The current approach to assessment of student achievement which relies on multiple choice student response must be abandoned because of its deleterious effect on the education process. An assessment system which measures student achievement on performance-based measures is essential for driving the needed reform toward a thinking curriculum in which students are actively engaged and successful in achieving goals in and beyond high school. (p. 17)

It is argued that new performance-based assessments can be designed to be so closely linked to the goals of instruction as to be almost indistinguishable from them. Rather than being a negative consequence, as it is now with some high-stakes uses of existing standardized tests, teaching to these proposed performance-based assessment would be considered a virtue.

---

[2] Debra P. v. Turlington, 644 F.2d 397, (5th Cir. 1981) at 404.

A wide range of approaches have been labeled performance assessments. Although there are no agreed-upon defining characteristics, the term generally refers to assessment tasks that require students to perform an activity (e.g., a laboratory experiment in science) or construct a response. Extended periods of time, ranging from several minutes to several weeks, may be needed to perform a task. Often the tasks are simulations of, or representations of, criterion activities valued in their own right. Evaluations of performance depend heavily on professional judgment.

As with many hot developments in education, performance-based assessments are being put forward as if they were new innovations in measurement. As Mehrens (1992) has noted, however, "performance assessment is not new" (p. 3). Various types of performance assessments were the norm before the introduction of multiple-choice testing in this country and remain the norm in many other countries. Performance assessment is also a part of the day-to-day classroom activities for many teachers.

Although the use of male pronouns and possibly the use of the phrase "criterion situations" rather than a term such as "authentic" or "valued, real-world" activities may date the quote, the words written by Lindquist (1951) more than 40 years ago would otherwise seem quite consistent with the current movement toward performance-based assessments. "... the most important consideration is that the test questions require the examinee to do the *same* things, *however complex*, that he is required to do in the criterion situations" (p. 154, emphasis in the original). Lindquist also recognized that, due to practical constraints, this ideal can only be approximated in practice. "It should always be the fundamental goal of the achievement test constructor to make the elements of his test series as nearly equivalent to, or as much like, the elements of the criterion series as consequences of efficiency, comparability, economy, and expediency permit" (p. 152).

Generalizability. Efficiency, comparability, and economy pose potentially formidable stumbling blocks for the implementation of a performance-based examination system of the type being proposed. As several authors (e.g., Dunbar, Koretz, & Hoover, 1991; Herman, 1991; Linn et al., 1991; Mehrens, 1992; Shavelson, Baxter, & Pine, 1992) have noted, one of the major stumbling blocks in this regard stems from the limited degree of generalizability of performance from one task to another.

Because the ratings of performance assessments depend upon professional judgments, there is a concern regarding the comparability of ratings assigned by different judges. Indeed, proponents of the "new objective tests" in the early 1900s used the argument that subjective judgments of student essays were inherently unfair to argue against performance-based assessments. The classic studies of Starch and Elliot (1912, 1913) conducted some 80 years ago, for example, demonstrated the extraordinary range of grades that teachers would assign to a single written essay or extended response in geometry. Although such demonstrations did much to discredit subjective scoring of open-ended responses and thereby to provide indirect support for the new objective tests then being championed, the approach was hardly a fair test of the use of professional judgment to score extended responses because it lacked any use of agreed-upon scoring rubrics or training of raters.

Judges contribute to the error variance of ratings of performance-based assessments. However, with careful design of scoring rubrics and training of raters, the magnitude of the variance components due to raters or interactions of raters with examinees can be kept at levels substantially smaller than other sources of error variance, the most notable of which is topic or task specificity (Baker, 1992; Dunbar et al., 1991; Shavelson et al., 1992). This finding is hardly new, but deserves renewed attention in light of the current emphasis on performance-based assessment.

Experience with performance-based measures in a variety of contexts indicates that performance is highly task-specific. That is, performance on one task has only a weak to modest relationship to performance on another, even seemingly similar, task. Ratings of student essays, for example, show considerable variability in performance as the consequence of the specific prompt or writing task (e.g., Breland, Camp, Jones, Morris, & Rock, 1987; Coffman, 1966; Dunbar et al., 1991; Hieronymus & Hoover, 1987). Similar variability in performance due to choice of task has been found with hands-on science tasks (Shavelson et al., 1992), with hands-on performance tasks for military jobs (Shavelson, Mayberry, Li, & Webb, 1990), and with performance tasks used in medical licensure examinations (Swanson, Norcini, & Grosso, 1987). The following examples illustrate the limited degree of generalizability across tasks that has been found in a variety of assessment contexts.

Figure 3 demonstrates the greater importance of increasing the number of tasks than increasing the number of trained raters for purposes of improving generalizability (using results from Baker's (1992) performance-based history tasks). Baker's history tasks require students to read original source documents such as the Lincoln-Douglas debate, the 1987 Brooks-Pixley debate on Chinese immigration, and the 1981 Simon-Graham debate on immigration. Students then write a 50-minute essay that draws on their prior knowledge of historical concepts and facts, together with the readings, to explain the most important ideas and issues. As can be seen, increasing the number of essays that each examinee is required to write yields substantially greater gains than can be achieved by increasing the number of raters.

A similar pattern of more rapidly increasing generalizability as a function of number of tasks rather than number of raters is shown in Figure 4 for open-ended mathematics problems reported by Suzanne Lane and her colleagues (Lane, Stone, Ankenmann, & Liu, 1992). The time allowed per problem was approximately 5 minutes. As might be expected with relatively short problems in mathematics, trained raters achieve a high degree of agreement so there is even less to be gained by increasing the number of raters per class for these tasks than for the more extended history tasks used by Baker.

Intercorrelations shown in Table 1 for hands-on science tasks developed by Shavelson and his colleagues (Shavelson, Baxter, Pine, & Yure, 1991) again show a high degree of task specificity in performance. As was true of Baker's history assessments and Lane's mathematics assessments, Shavelson et al. (1991) found high levels of interrater reliability but limited generalizability across tasks. Results such as those in Table 1 led Shavelson et al. (1992) to conclude that "some students performed well on one task (e.g., mystery box or bug experiment) while other students performed well on another task" (p. 25).

Experiences with performance-based licensure examinations in law and medicine confirm the need for a large number of tasks, which translates into a substantial number of hours of testing, in order to achieve acceptable levels of generalizability. Figure 5 displays the level of generalizability as a function of hours of testing for the two open-ended problem sections of the California Bar Exam based on results reported by Klein (personal communication, November 14, 1991). Examinees are given reference materials and allowed three hours to

Figure 3

Score Generalizability of General Impression Content Quality Scores of Extended History Tasks as a Function of Number of History Topics and Number of Raters (Based on Baker, 1992)

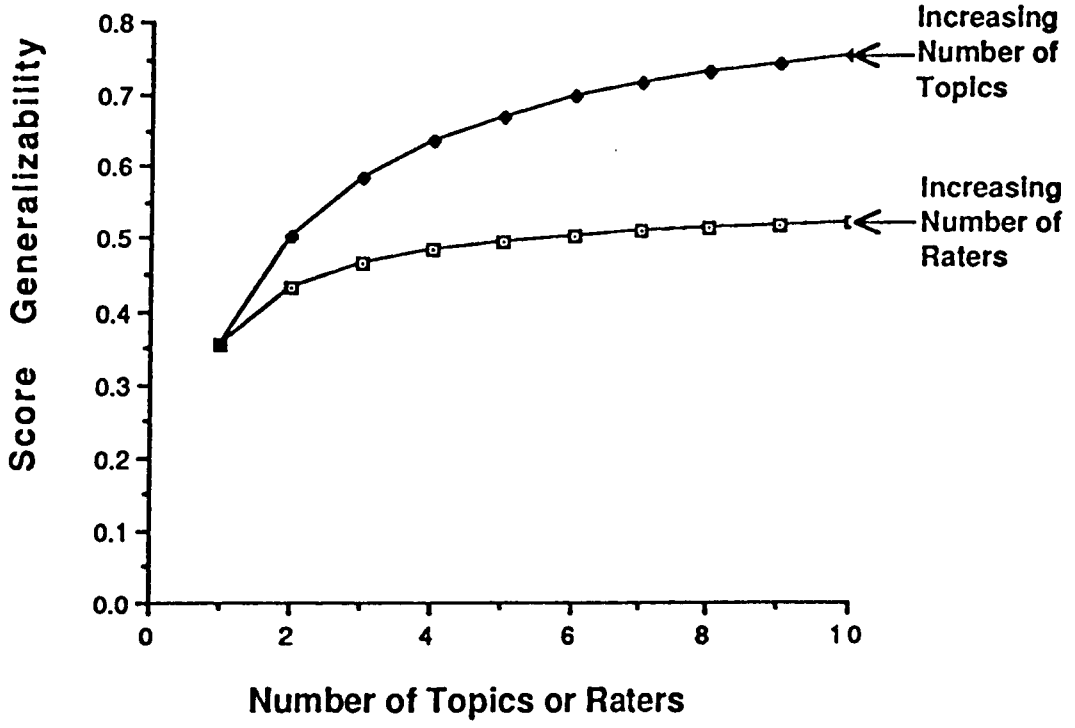


Figure 4

Score Generalizability of QUASAR Form D Mathematics Scores Estimated from Rater Pair 13 (Based on Lane, et al., 1992)

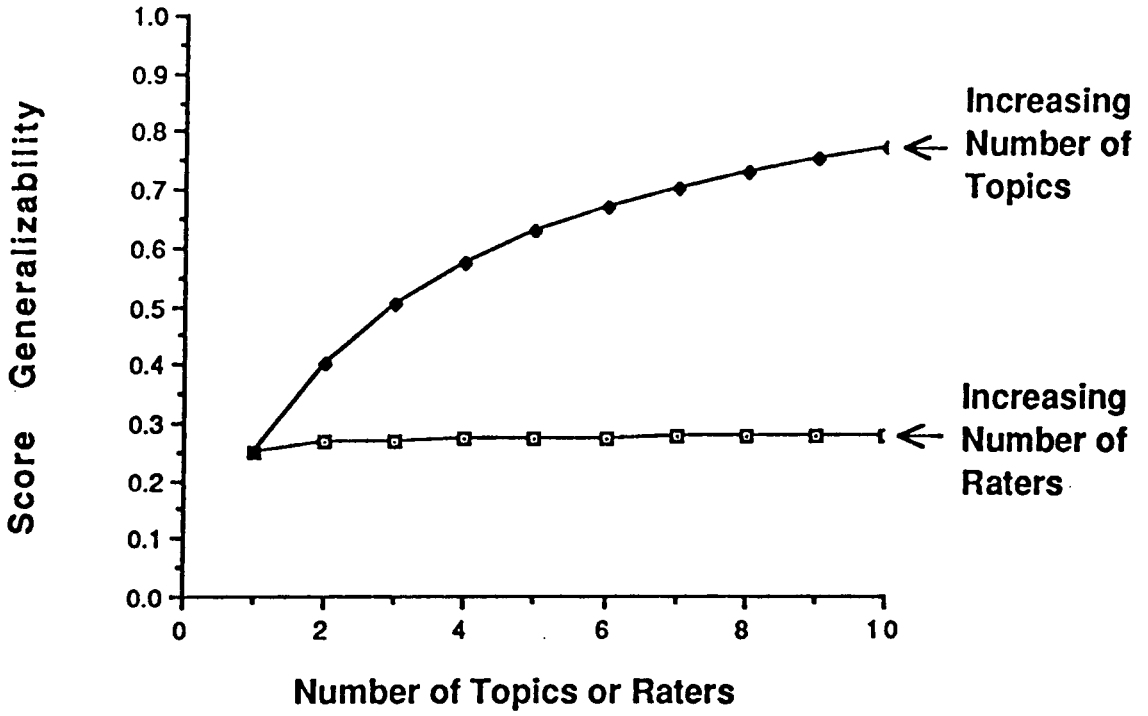


Table 1

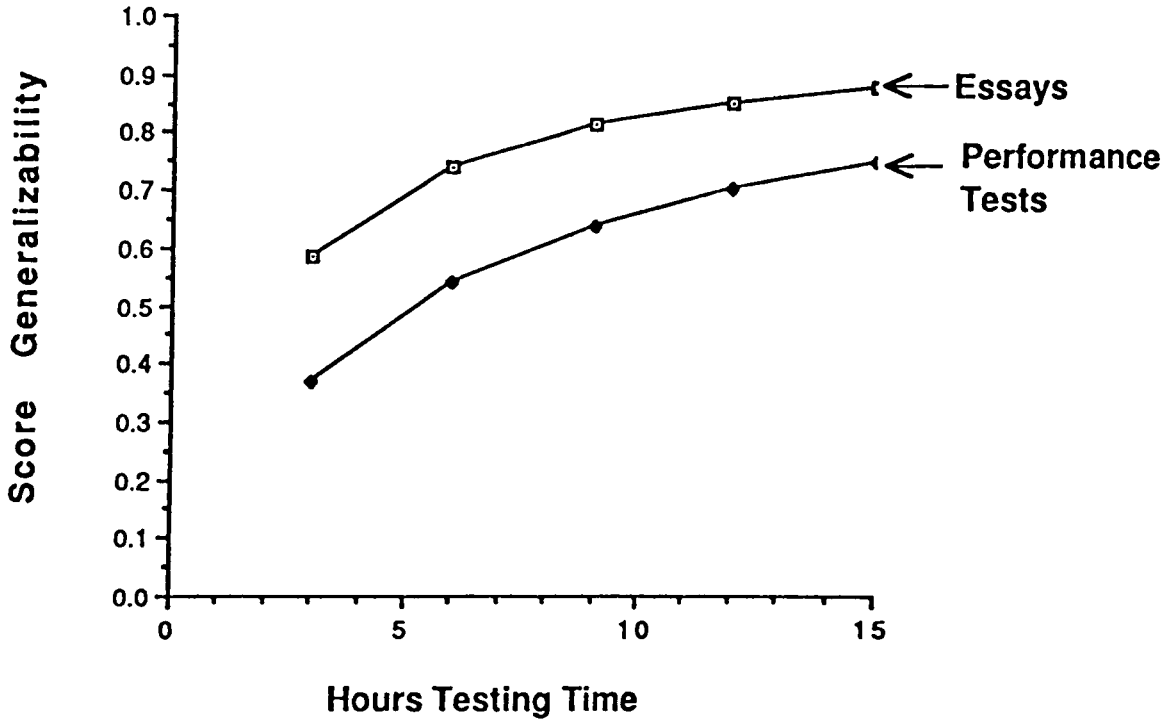
Intercorrelations of Three Hands-On Science Tasks (Based on Shavelson, Baxter, Pine, & Yure, 1990)*

Task	Paper Towels	Electronic Mysteries	Sow Bugs
Paper Towels	---	.14	.24
Electronic Mysteries	.25	---	.39
Sow Bugs	.40	.35	---

* District A (traditional instruction) below diagonal and District B (hands-on instruction) above diagonal.

Figure 5

California Bar Exam Essay and Performance Test Generalizability Coefficients as a Function of Testing Time (Based on Klein, 1991)



complete each performance test. Since the inter-task correlations of the performance tests are only slightly higher than the correlations between scores on the one-hour essays, it would require approximately two and a half times as long to achieve a given level of generalizability using performance tests as it would with essay tests.

Van der Vleuten and Swanson (1990) reviewed the generalizability results from nine different studies investigating the assessment of clinical skills using standard patients. A summary of their results is shown in Figure 6. Although there is considerable between-study variability in the level of generalizability for a given length of testing time, a substantial amount of testing time is needed to achieve a reasonable degree of generalizability in all cases.

My final generalizability examples are based on results from Advanced Placement (AP) examinations (College Board, 1988), which have for some time used a combination of multiple-choice items and performance-based tasks. The nature of the performance-based tasks varies substantially in degree of structure and the amount of time allowed per task. In Music Theory, for example, examinees have an hour for 6 open-ended problems, while in American History they have an hour and 45 minutes for 2 problems. As can be seen in Figure 7, the typical inter-task correlations also vary substantially across subjects, ranging from .20 for the History of Art examination to .67 for the Physics C: Electricity and Magnetism examination.

Figure 6

Score Generalizability of Assessments of Clinical Skills with Standard Patients as a Function of Testing Time (Based on van der Vleuten & Swanson, 1990)

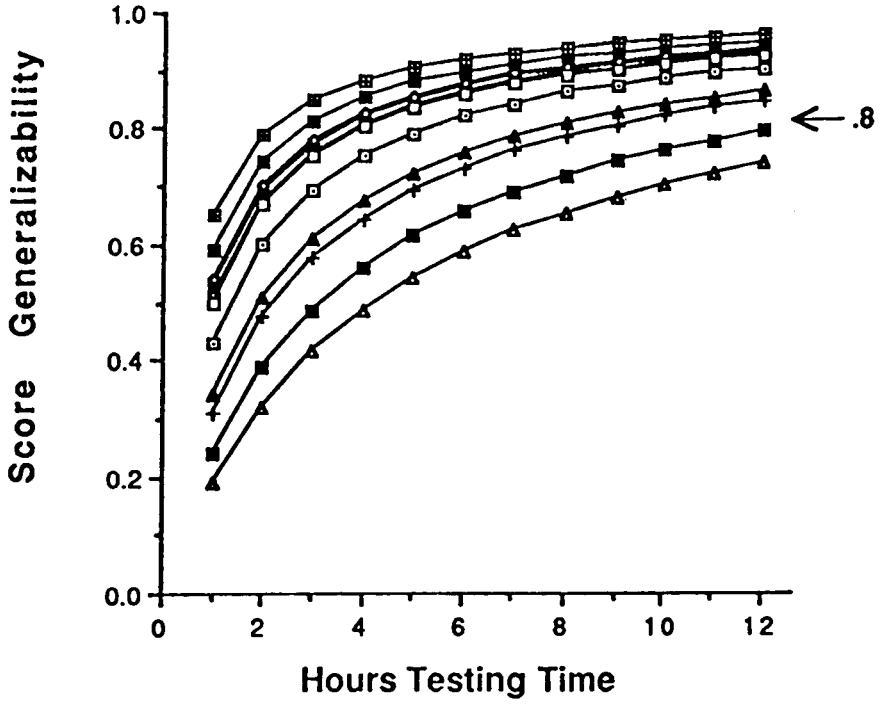
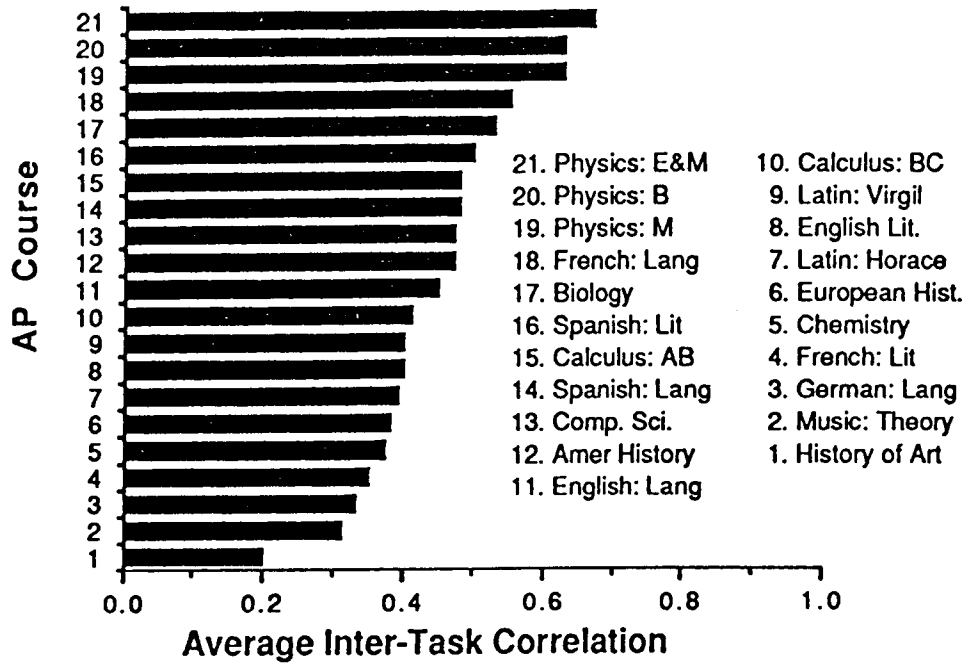


Figure 7

Average Inter-Task Correlations for the Free-Response Sections of Advanced Placement Examinations for 21 Courses*



* Derived from data reported for 1986 test administrations in Advanced Placement Technical Manual, College Board, 1988, pp. 68-95)

As the scatterplot of average inter-task correlation with average number of minutes per task in Figure 8 shows, there is no relationship between the amount of time allowed to complete tasks on AP examinations in different subjects and the degree of inter-task correlation. This lack of relationship, coupled with the subject-to-subject variability in time per task and average inter-task correlations, leads to an even greater variability in the amount of testing time that would be required to achieve a generalizability coefficient of .90 or higher if the AP exams consisted of only the current types of performance-based problems. The estimated required amounts of testing time shown in Figure 9 range from a low of an hour and 15 minutes for Physics C: Electricity and Magnetism to 13 hours for European History. Six or more hours would be required for 8 of the 21 subjects.

Mehrens (1992) recently argued that "the major problems for valid performance assessment relate to the limited sampling and lack of generalizability from the limited sample to any identifiable domain" (p. 7). The results that have just been summarized provide support for that conclusion. High levels of generalizability across tasks alone *do not guarantee* valid inferences or justify particular uses of assessment results. But low levels of generalizability across tasks limit the validity of inferences about performance for a domain and pose serious problems regarding comparability and fairness to individuals who are judged against performance standards based on too small a number, and perhaps a different set, of tasks.

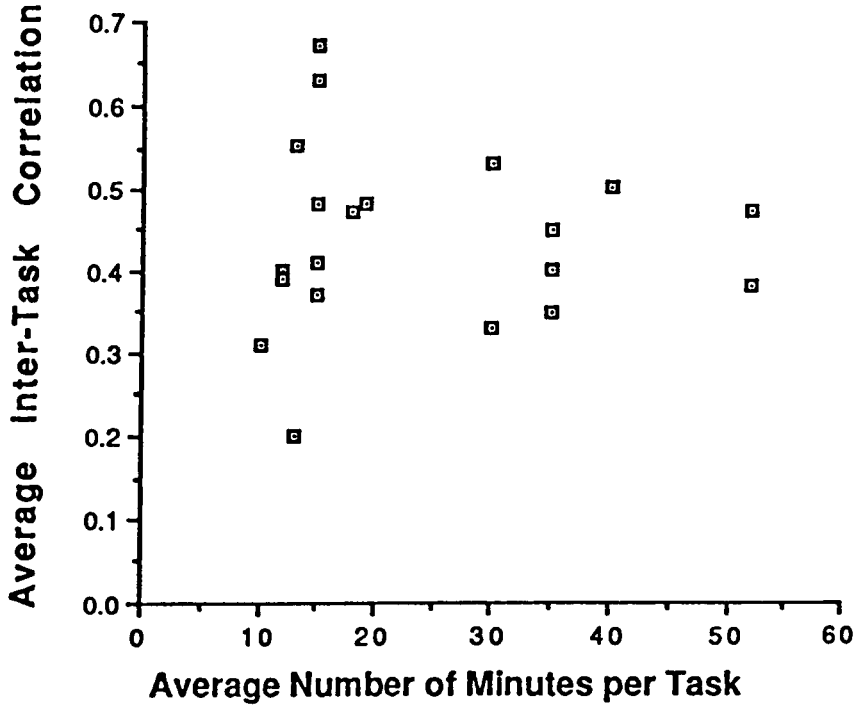
Implications for Proposed Assessment Systems

The primary way of dealing with the lack of generalizability across tasks is to increase the number of tasks on the assessment. Including a relatively large number of tasks is a simple solution where each task requires little time. A requirement of a large number of tasks obviously poses more of a problem and requires more detailed justification, however, when respondents need a substantial period of time to complete each task.

Substantial numbers of more time-consuming tasks can be justified in at least two ways. First, in high-cost or high-risk situations such as the licensing of a doctor, the time and expense to achieve more valid measurement can readily be justified. For example, if computer-based patient simulation problems are judged to provide a more valid basis for licensing physicians,

Figure 8

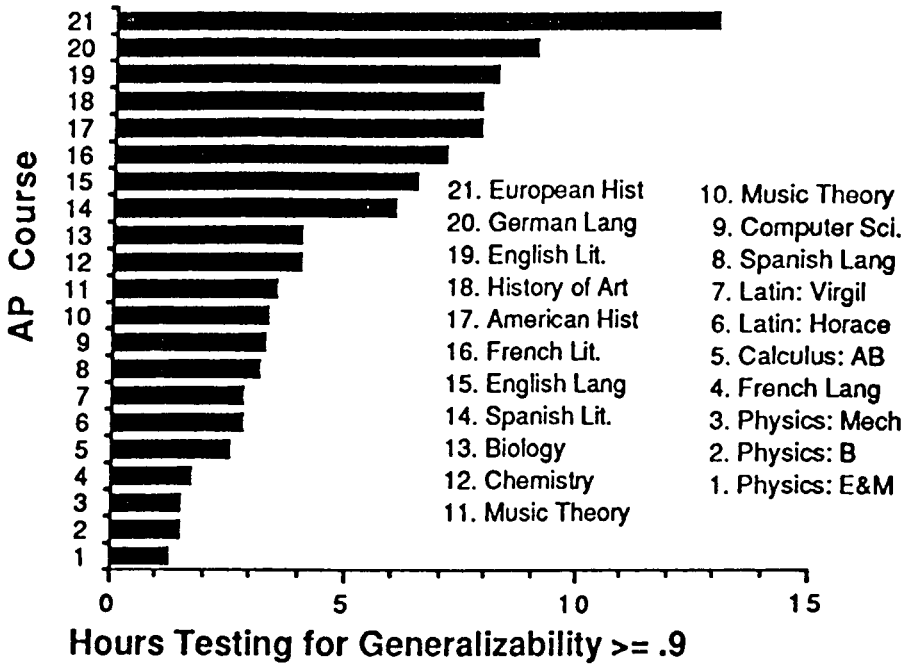
Plot of Average Inter-Task Correlations vs. Average Number of Minutes Per Task for Free-Response Sections of Advanced Placement Examinations for 21 Courses*



* Derived from data reported for 1986 test administrations in Advanced Placement Technical Manual, College Board, 1988, pp. 68-95)

Figure 9

Estimated Hours of Testing Time Needed for a Generalizability Coefficient of .90 or Higher on the Free-Response Sections of Advanced Placement Examinations for 21 Courses*



* Derived from data reported for 1986 test administrations in Advanced Placement Technical Manual, College Board, 1988, pp. 68-95)

then such an approach can be justified despite the fact that Julian and Wright (1988) found that a minimum of eight problems, each requiring approximately an hour and a half to complete, would be needed to achieve an acceptable level of generalizability.

A second possible justification for the inclusion of multiple tasks that require substantial amounts of time is that task performance is itself a beneficial part of instruction. That is, the tasks provide useful learning experiences as well as information about a student's current level of competency. This second justification is likely to be more important in the case of the assessments proposed by NCEST and by SCANS, which are envisioned as being closely integrated with instruction. Assessments that are an integral part of instruction require that the tasks are valued learning activities in their own right. This goal will be an important consideration in the design and evaluation of tasks for the proposed national systems of examinations.

The premise that proposed high-stakes examination systems with heavy, or possibly exclusive, reliance on performance-based assessments will have beneficial effects also underscores the need to emphasize the evaluation of the consequences of the system. Dunbar et al. (1991) recently observed that "the nation stands poised on the brink of yet another wave of test-based reform, and again we appear prepared to undertake it without sufficient quality control" (p. 302). The quality control that they argue for would include investigations that would address both the evidential and consequential bases for valid interpretation and use of assessment results that Messick (1989) has articulated. It is incumbent upon the measurement research community to make the case that the introduction of any new high-stakes examination system include provisions for paying greater attention to investigations of both the intended and unintended consequences of the system than has been typical of previous test-based reform efforts. As Messick (1992) recently noted, "This evidence should especially address both the anticipated consequences of performance assessment for teaching and learning as well as potential adverse consequences bearing on issues of bias and fairness" (p. 35 of typescript).

References

Baker, E. L. (1992). *The role of domain specifications in improving the technical quality of performance assessment* (CSE Tech. Rep.). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Berliner, D. C. (1992, February). *Educational reform in an era of disinformation*. Paper presented at the annual meeting of the American Colleges of Teacher Education, San Antonio, TX.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill* (Research Monograph No. 11) New York: College Entrance Examination Board.

California State Department of Education. (1990). *Education Summit*. Sacramento, CA: California State Department of Education.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends for Education.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice, 7*(2), 5-9.

Carson, C. C., Huelskamp, R. M., & Woodall, T. D. (1991). *Perspective on education in America* (3rd draft). Albuquerque, NM: Scandia National Laboratories.

Catterall, J. (1987). *Standards and school dropouts: A national study of minimum competency testing* (CSE Tech. Rep. 278). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3,* 151-156.

College Board. (1988). *Technical manual for the Advanced Placement program*. New York: College Board.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity*, (pp. 3-17). Hillsdale, NJ: Erlbaum.

Dunbar, S. B., Koretz, D., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 289-304.

Herman, J. (1991). Research in cognition and learning: Implications for achievement testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 154-165). Englewood Cliffs, NJ: Prentice Hall.

Hieronymus, A. N., & Hoover, H. D. (1987). *Iowa Tests of Basic Skills: Writing supplement teacher's guide.* Chicago: Riverside.

Jaeger, R. M. (1992, April). *"World class" standards, choice, and privatization: Weak measurement serving presumptive policy*. Division D, Vice Presidential Address presented at the annual meeting of the American Educational Research Association, San Francisco.

Julian, E. R., & Wright, B. D. (1988). Using computerized patient simulations to measure the clinical competence of physicians. *Applied Measurement in Education, 1*, 299-318.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *Effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1992, April). *Empirical evidence for the reliability and validity of performance assessments*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 119-158). Washington, DC: American Council on Education.

Linn, R. L. (1987). Accountability: The comparison of education systems and the quality of test results. *Educational Policy, 1*, 181-198.

Linn, R. L. (in press). Educational reform through national standards and choice: An analysis of underlying premises. *Educational Policy*.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5-14.

Madaus, G. F. (1985). Public policy and the testing profession: You've never had it so good? *Educational Measurement: Issues and Practice, 4*(4), 5-11.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3-9, 20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York: Macmillan.

Messick, S. (1992, April). *The interplay between evidence and consequences in the validation of performance assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Mullis, I. V. S., Dossey, J. A., Foertsch, M. A., Jones, L. R., & Gentile, C. A. (1991). *Trends in academic progress: Achievement of U.S. students in science, 1969 to 1990; mathematics, 1973 to 1990; reading, 1971 to 1990; writing, 1984 to 1990* (Report No. 21-Y-01; ISBN 0-16-036046-3). Washington, DC: U.S. Department of Education.

National Council on Education Standards and Testing (NCEST). (1992). *Raising standards for American education*. Washington, DC: U.S. Government Printing Office.

Petrie, H. G. (1987). Introduction to "evaluation and testing." *Educational Policy, 1*, 175-180.

Resnick, L. B. (1990, October). *Assessment and educational standards.* Paper presented at the Office of Educational Research and Improvement conference, The Promise and Peril of Alternative Assessment, Washington, DC.

Resnick, L. B. (1992). Why we need national standards and exams. *State Education Leader*, *11*(1), 4–5. Denver, CO: Education Commission of the States.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 9-350. Boston, Kluwer Academic Publishers.

Shavelson, R. J., Baxter, G. P., & Pine J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shavelson, R. J., Baxter, G. P., Pine, G. P., & Yure, J. (1991, April). *Alternative assessment technologies for assessing science understanding.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Shavelson, R. J., Mayberry, P., & Li, W., & Webb, N. M. (1990). Generalizability of job performance measurements: Marine Corps rifleman. *Military Psychology, 2*, 129-144.

Smith, M. S., O'Day, J., & Cohen, D. K. (1990). National curriculum American style: Can it be done? What might it look like? *American Educator, Winter,* 10-17, 40-47.

Starch, D., & Elliot, E. C. (1912). Reliability of grading high school work in English. *School Review, 20,* 442-457.

Starch, D., & Elliot, E. C. (1913). Reliability of grading high school work in mathematics. *School Review, 21,* 254-259.

Swanson, D., Norcini, J., & Grosso, L. (1987). Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12,* 220-246.

U. S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

U. S. Department of Education. (1991) *America 2000: An education strategy.* Washington, DC: U. S. Department of Education.

U. S. Department of Labor, The Secretary's Commission on Achieving Necessary Skills. (1991). *What work requires of schools: A SCANS report for America 2000.* Washington, DC: U.S. Department of Labor.

van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2,* 58-76.

Wainer, H. (1986). The SAT as a social indicator: A pretty bad idea. In *Drawing inferences from self-selected samples* (pp. 7-21). New York: Springer-Verlag.