DOCUMENT RESUME

ED 390 891 TM 024 172

AUTHOR Messick, Samuel
TITLE The Interplay of Evidence and Consequences in the Validation of Performance Assessments. Research Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-39
PUB DATE Jul 92
NOTE 47p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cognitive Processes; *Competence; *Educational Assessment; *Evaluation Methods; Performance; Test Items; Test Results; *Test Validity
IDENTIFIERS Authentic Assessment; Direct Assessment; *Evidence; *Performance Based Evaluation

ABSTRACT

Authentic and direct assessments of performances and products are conceptualized in terms of multiple distinctions having implications for validation. These include contrasts between performances and products, between assessment of performance per se and performance assessment of competence or other constructs, between structured and unstructured problems and response modes, between decomposed task skills and complex task performance, and between contextualized and decontextualized knowledge and skill. The concepts of "authenticity" and "directness" of assessment are analyzed as promissory validity claims that they offset, respectively, the two major threats to construct validity, namely, construct underrepresentation and construct-irrelevant variance. These distinctions are examined in the context of an overarching contrast between task-driven and construct-driven performance assessment. With respect to validation, the salient role of both positive and negative consequences is underscored as well as the need for evidence bearing on the various aspects of construct validity (content, substantive, structural, external, generalizability, and consequential). (Contains 51 references.) (Author)

RESEARCH REPORT

THE INTERPLAY OF EVIDENCE AND CONSEQUENCES IN THE VALIDATION OF PERFORMANCE ASSESSMENTS

Samuel Messick

Educational Testing Service
Princeton, New Jersey
July 1992

THE INTERPLAY OF EVIDENCE AND CONSEQUENCES IN THE VALIDATION OF PERFORMANCE ASSESSMENTS

Samuel Messick
Educational Testing Service

THE INTERPLAY OF EVIDENCE AND CONSEQUENCES IN THE VALIDATION OF PERFORMANCE ASSESSMENTS

Samuel Messick
Educational Testing Service

Abstract

Authentic and direct assessments of performances and products are conceptualized in terms of multiple distinctions having implications for validation. These include contrasts between performances and products, between assessment of performance per se and performance assessment of competence or other constructs, between structured and unstructured problems and response modes, between decomposed task skills and complex task performance, and between contextualized and decontextualized knowledge and skill. The concepts of "authenticity" and "directness" of assessment are analyzed as promissory validity claims that they offset, respectively, the two major threats to construct validity, namely, construct underrepresentation and construct-irrelevant variance. These distinctions are examined in the context of an overarching contrast between task-driven and construct-driven performance assessment. With respect to validation, the salient role of both positive and negative consequences is underscored as well as the need for evidence bearing on the various aspects of construct validity (content, substantive, structural, external, generalizability, and consequential).

THE INTERPLAY OF EVIDENCE AND CONSEQUENCES IN THE VALIDATION OF PERFORMANCE ASSESSMENTS

Samuel Messick[1]
Educational Testing Service

Performance assessments are becoming increasingly popular because they promise authentic and direct appraisals of educational competence leading to positive consequences for teaching and learning. These positive consequences involve, among other things, the expectation of enhanced student learning not just of basic skills, but also of higher-order thinking skills such as problem representation, reasoning, judgment, and synthesis. Assessments of performance, by virtue of focussing not just on what students know but on what they do and the way they do it, also promise to facilitate the process of learning-by-doing as well as the development of generic skills for written and oral communication, executive planning, interpersonal acumen, and other enabling competencies.

However, along with the promised benefits, there also come problems of implementation and validation. Furthermore, the problems as well as the benefits of performance measurement redound both to its use in instructional assessment, which serves to facilitate and monitor day-to-day teaching and learning (Nitko, 1989; Wiggins, 1989), and in accountability assessment, which serves to certify demonstrated competence at whatever level of aggregation (Mehrens, 1992). Although the present paper focusses on validity issues in the certification of competence, many of its points apply as well to instructional assessment. However, because the stakes are lower and decision

---

[1]Acknowledgments are gratefully extended to Isaac Bejar, Randy Bennett, Robert Linn, and Warren Willingham for their reviews of the manuscript; to Richard Snow for suggesting a number of ways in which points could be amplified or strengthened; and, to Ann Jungeblut for helping to clarify both the thinking and the writing.

This paper was delivered as an invited address to the annual meeting of the National Council on Measurement in Education, San Francisco, April, 1992.

errors are easier to recover from in this latter case, some -- but not all -- of the technical standards may be relaxed somewhat in instructional assessment (Messick, 1992a).

The portrayal of performance assessments as being authentic and direct has all the earmarks of a validity claim but with little or no evidential grounding. We need to address what the labels "authentic" and "direct" might mean in validity terms. We also need to determine what kinds of evidence might legitimize both their usage as validity standards and their nefarious implication that other forms of assessment are not only indirect, but inauthentic.

With respect to consequences as validity evidence, I have argued for nearly 30 years that test validity and social values are intertwined and that evaluation of intended and unintended consequences of any testing is integral to the validation of test interpretation and use (Messick, 1964, 1965, 1975, 1981, 1988, 1989). However, until the recent upsurge of renewed interest in performance assessment, there have been relatively few adherents to this position among measurement practitioners. Because they are now singing an old favorite song, the refrain of which intones that the consequences of measurement betoken its validity, I confess a certain fondness for performance assessors. But at the same time I am concerned that their enthusiastic embracing of the consequential basis of test validity might lead to a shortchanging of the evidential basis, including the need for evidence of the consequences.

Because "there is no absolute distinction between performance tests and other classes of tests" (Fitzpatrick & Morrison, 1971, p. 238), performance assessments must be evaluated by the same validity criteria, both evidential and consequential, as are other assessments. Different psychometric models might be employed, to be sure, as well as different scoring procedures and rubrics, but such basic assessment issues as validity, reliability, comparability, and fairness still need to be uniformly addressed. This is so because validity, reliability, comparability, and fairness are not just measurement principles, they are *social values* that have meaning and force outside of measurement wherever evaluative judgments and decisions are made.

Hence performance assessments, like all assessments, need to be responsive to important and persistent validity questions, such as:

> Are we looking at the right things in the right balance?
>
> Has anything important been left out?
>
> Does our way of looking introduce sources of invalidity or irrelevant variance that bias the scores or judgments?
>
> Does our way of scoring reflect the manner in which domain processes combine to produce effects and is our score structure consistent with the structure of the domain about which inferences are to be drawn or predictions made?
>
> What evidence is there that our scores mean what we interpret them to mean, in particular, as reflections of knowledge and skill having plausible implications for educational action relative to personal or group standards?
>
> Are there plausible rival interpretations of score meaning or alternative implications for action and, if so, by what evidence and arguments are they discounted?
>
> Are the judgments or scores reliable and are their properties and relationships generalizable across the contents and contexts of use as well as across pertinent population groups?
>
> Do the scores have utility for the proposed purposes in the applied settings?
>
> Are the scores applied fairly for these purposes?
>
> Are the short- and long-term consequences of score interpretation and use supportive of the general testing aims and are there any adverse side-effects?

Which, if any, of these questions is unnecessary to address in justifying score interpretation and use? In any event, this list of questions is intended to be stimulative, not exhaustive.

In various ways, these questions address the six aspects of unified validity delineated in the third edition of *Educational Measurement* (Linn, 1989), that is, content, substantive, structural, external, generalizability, and consequential aspects of construct validity (Messick, 1989). In brief, the content aspect of construct validity includes evidence of content

relevance, representativeness, and technical quality; the substantive aspect refers to theoretical rationales for consistencies in test responses, including process models of task performance; the structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue; the external aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons as well as evidence of criterion relevance and applied utility; the generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks, including validity generalization of test-criterion relationships; and, the consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of test invalidity related to issues of bias, fairness, and distributive justice (Messick, 1989, 1992b).

The general thrust of such questions and of the validity aspects underlying them is to seek evidence and arguments to discount the two major threats to construct validity -- namely, construct underrepresentation and construct-irrelevant variance -- as well as to evaluate the action implications of score meaning. In effect, the six interrelated aspects of unified validity provide a general framework for the validation of all assessments including performance assessments. Validity criteria specialized for performance assessments -- as proposed, for example, by Linn, Baker, and Dunbar (1991) and in a more limited way by J. R. Frederiksen and Collins (1989) -- are for the most part consistent with this general framework. Specifically, Linn and his colleagues propose content quality, content coverage, cognitive complexity, meaningfulness, cost and efficiency, transfer and generalizability, fairness, and consequences as specialized validity criteria; J. R. Frederiksen and Collins propose directness, scope, reliability, and transparency. However, to the degree that performance assessments are evaluated in terms of the full range of general criteria (with cost-utility not being preemptive but, rather, only one of several standards contributing to a balanced judgment), then special validity dispensations are not needed to legitimize the performance-based approach, thereby making its scientific foundation and practical credibility that much stronger.

This is not a major issue in principle because the specialized validity criteria, especially as proposed by Linn and his colleagues, fit well with and usefully elaborate the general criteria. However, it may become an issue in validation practice because these specialized criteria, although not intended to be exhaustive, are nonetheless less extensive. If exclusively relied upon, some validity aspects might be downplayed or left out, particularly those bearing on score interpretation and its value implications. Validation criteria will be examined later after drawing some conceptual distinctions in connection with the functions, types, and purposes of performance assessment that have implications for validation.

Although on the surface these distinctions are not new ones, the intent is to test their limits and explore their intended and unintended implications for the development, interpretation, use, and validation of performance assessments. For example, they include contrasts between performances and products, between assessment of performance and performance assessment of competence or other constructs, between structured and unstructured problems and response modes, between decomposed task skills and complex task performances, between contextualized and decontextualized knowledge and skill, between authentic and inauthentic tasks and assessments, and between direct and indirect appraisal of behavior and constructs. These complex contrasts are often cast as simple dichotomies and sometimes amount to false dichotomies. It is important to probe their deep structure because the imposition of dichotomies on continua or multidimensional arrays unnecessarily polarizes ideological orientations toward the uses of assessment and severely reduces available options for the improvement and diversification of measurement. These contrasts are examined in the context of an overarching distinction having implications fundamental to many of the other points being made, namely, the contrast between task-centered and competency- or construct-centered performance assessment.

TASK-DRIVEN VERSUS CONSTRUCT-DRIVEN PERFORMANCE ASSESSMENT

Before exploring the differential implications of focussing on tasks as opposed to constructs in the development and interpretation of performance

assessments, let us first examine the distinction between performances and products because, although different, they are both ordinarily viewed as equally important exemplars of performance assessment. Indeed, the term "performance assessment" traditionally serves as a convenient shorthand for the more complete "performance-and-product assessment" (Fitzpatrick & Morrison, 1971). In this context, we also contrast the assessment of performances and products per se with their use as vehicles for the assessment of competencies or other constructs.

Performances and Products as Targets and Vehicles of Assessment

In the rush to move performance assessment forward, one often gets the impression that any product or performance or student-constructed answer substituted for (or even added to) multiple-choice items would serve the cause equally well. However, a more principled approach is called for. There should be a guiding rationale akin to test specifications that ties the assessment of particular products or performances to the purposes of the testing, to the nature of the substantive domain at issue, and to construct theories of pertinent skills and knowledge.

In the first instance, the decision to assess either a performance or a product depends on the educational objective that the measurement serves to monitor or advance. For example, if an educational focus in science teaching is on laboratory skills, then assessing knowledge of experimental techniques by means of an essay as opposed to an objective test is not an authentic substitute for appraising the conduct of an experiment. In addition, there are a number of other considerations affecting the choice of performance as opposed to product as a basis for assessment, such as the nature of the substantive domain and the differential roles of performance and product creation within it (Fitzpatrick & Morrison, 1971).

In some domains such as acting and dancing, the performance and the product are essentially the same thing, differing mainly in whether it is viewed live or recorded in some form. Information relevant to the assessment is in the way in which the acting or dancing is carried out, with little or no difference in pertinent information revealed in the live performance or the recorded product. Thus, in domains like the performing arts, all there is to

evaluate are performances in one form or another. To the extent that it is also desirable to assess the knowledge base for the performance, for example, or knowledge about techniques of performance, then other products or performances as well as other modes of assessment might be called for.

In other domains such as painting or creative writing, there may be so many acceptable variations in process or alternative modes of proceeding that the outcome or product is all that counts. Information relevant to the assessment is contained in the product, at least in good products, with little or nothing added by examining the processes of production. To the extent that the processes of production do contain information relevant to fostering growth or providing critical feedback, of course, then they too should be evaluated, as in instructional uses of portfolios (Wolf, Bixby. Glenn, & Gardner, 1991). In addition, the knowledge base for the production might be assessed via student reflection on product development or on portfolio content or by essays or even objective tests. In domains where performance is intensely knowledge-based, such as teaching, a variety of products (lesson plans, critiques of student work, and so forth) as well as structured appraisals of knowledge via tests or interviews would likely need to accompany performance observations to buttress interpretations of teaching competence.

In still other domains such as auto mechanics and chemical experimentation, however, both the performance and the product warrant scrutiny from the outset, because not only is the outcome at issue but so are proper procedures, for example, to avoid deleterious side effects such as accidents. In general, one should consider assessing performances if task procedures have been explicitly taught and deviations from accepted practice can be detected, whereas assessment of products should be considered if proper task procedures are diverse or indeterminate or have not been explicitly taught (Fitzpatrick & Morrison, 1971).

Under certain circumstances, evaluation of the product or performance per se is the focus, that is, they are the targets as opposed to the vehicles of assessment. This might occur, for example, in an arts contest or a figure-skating competition or a science fair. In such cases, replicability and generalizability are not at issue. It does not matter how good a candidate's other recitals, skating programs, or science projects were or will be. All

that counts is the quality of the performance or product submitted for evaluation, and the validation focus is on the judgment of quality. But note that in this usage of performance assessment, inferences are not to be made about the competencies or other attributes of the performers, that is, inferences from observed behavior to constructs such as knowledge and skill underlying that behavior.

Such inferences about competencies or personal attributes require convergent and discriminant evidence to support the construct interpretation of performance scores and to discount plausible rival interpretations. The validation focus is on score meaning and its action or value implications as well as on the potential and actual consequences of implementing score-based actions. In this regard, the term "score" is used generically here in its broadest sense to mean any coding or summarization, whether quantitative or qualitative, of observed consistencies or performance regularities on a test, questionnaire, observation procedure, or other assessment device such as work samples, portfolios, or realistic problem simulations.

In the performance assessment of competencies or other constructs -- that is, where the performance is the vehicle not the target of assessment -- replicability and generalizability can no longer be ignored. This is the case because the consistency or variability of the performances contributes to score meaning, as does generalizability from the sample of observed tasks to the universe of tasks relevant to the knowledge or skill domain at issue. In essence, replicability and generalizability establish boundaries on the meaning of the scores and on how consistent that meaning is likely to be. In effect, the meaning of the construct is tied to the range of tasks and situations that it generalizes and transfers to. The explicit inclusion of transfer here extends the construct's range of reference, as well as the needed supportive evidence, beyond the generalizability of performance to include the transferability or facilitation of related learning, which also extends the range of potential action implications of score meaning for educational purposes.

Problems arise when measurement practitioners try to have it both ways. That is, they focus on particular products or performances as if these were the targets of assessment, treating issues of domain coverage and

generalizability with *belle indifference*. At the same time, they use construct language to infer score meaning and associated action implications, with little or no attention to convergent and discriminant evidence needed to sustain that meaning. This might be defensible if the products or performances that are viewed as targets of assessment are actually targets of instruction. But we must ask ourselves how many educational objectives worthy of time and effort can be captured in a single task or a small set of tasks (or products or performances). Then compare this with the number of worthy objectives implying consistency in performance across varied tasks. In any event, even in cases where the particular product or performance is indeed an instructional target, the issues of domain coverage and generalizability may be moot, but not the evidential basis of construct inferences about knowledge, skill, or other attributes of the performer.

Content Coverage and Construct Generalizability in Performance Assessment

For many if not most advocates of performance assessment, the issues of domain coverage and generalizability are not contentious, but problematic. For example, Linn, Baker, and Dunbar (1991) list as validity criteria specialized for performance assessment not only ***content coverage*** but ***content quality***, as well as ***transfer and generalizability***. In regard to content coverage, they give primacy to the need for appropriate sampling of task processes in addition to traditional coverage of domain content. As in the delineation of general validity criteria highlighting content and substantive aspects of construct validity, this implies a coordinated need to move beyond traditional professional judgment of content to accrue construct-related evidence that the ostensibly sampled processes are actually engaged by respondents in task performance (Messick, 1989). Thus, the issue of domain coverage refers not just to the content representativeness of the construct measure but also to the process representation of the construct and the degree to which those processes are reflected in construct measurement (Embretson, 1983). This notion of construct representation and underrepresentation will resurface subsequently in connection with the meaning of authenticity in measurement.

As another instance, Wiggins (1989) maintains that "multiple and varied tests are required. In performance-based areas we do not assess competence on the basis of one performance. . . . Over time and in the context of numerous performances, we observe the *patterns* of success and failure and the reasons behind them" (p. 705). This may be feasible in instructional assessment or in appraisals of extensive portfolios but, under ordinary conditions of accountability assessment, trade-offs may be required between breadth of content coverage and the depth of process understanding promised by the use of extended performance tasks.

On the other hand, J. R. Frederiksen and Collins (1989) appear to be more tentative with respect to domain coverage and are essentially silent with respect to generalizability. In the testing system they envision, "the tests should consist of a representative set of tasks that cover the spectrum of knowledge, skills, and strategies needed for the activity or domain being tested" (p. 30). Yet, the validity standard they call ***scope*** indicates that "the test should cover, as far as possible, all the knowledge, skills, and strategies required to do well in the activity," with no mention of the domain (p.30). Does this mean that domain coverage is potentially achieved by means of a set of tests, but each single test includes tasks covering a particular domain activity? And what about the meaning of the test scores?

Does score meaning refer to behavioral task-based constructs such as criticizing a sonnet or deriving a general equation for conic sections? Or does it extend to knowledge-and-skill constructs such as poetic appreciation or analytic geometry or to still broader domains such as literary criticism or quantitative reasoning? Contrariwise, is the definition of activity or even domain limited to the array of knowledge, skills, and strategies that the test can feasibly cover? Or do the construct boundaries extend to the range of tasks and situations that the performance scores generalize and transfer to? It would seem that scope, as a validity standard, addresses what the test covers but not what the scores mean. Setting the boundaries of score meaning is precisely what generalizability evidence is meant to address but it finds no place in this or any of the other validity standards proposed by J. R. Frederiksen and Collins -- even their ***reliability*** standard, which refers only to reliability of scoring. However, they do urge students to repeat the test

multiple times to encourage striving for improvement, but this is not in the context of generalizability, even across the multiple testing occasions.

In regard to trade-offs between breadth and depth in coping with domain coverage and the limits of generalizability, Linn and his colleagues (1991) suggested that one could increase the number of performance assessments for each student or, presumably, increase the number of tasks in each assessment. Here the trade-off is between breadth of coverage and nonassessment instructional activities that might instead have filled the extended testing time. One way of justifying this might be to argue, with evidence, that the assessment experience is of comparable or higher educational value than the replaced instructional activities. Another suggestion was to use a matrix-sampling design with different performance tasks administered to different samples of students. Here the gain in breadth of coverage comes at the expense of individual student scores or, at least, of comparable individual scores (except in unidimensional or hierarchically structured domains where score comparability might be attained through complex sampling plans and strong psychometric models like item response theory). Nonetheless, matrix sampling is especially useful when the accountability concern focusses on some aggregate level such as the school, district, state, or nation.

Trade-Offs Between Structured and Open-Ended Assessments

Another approach for enhancing domain coverage and for appraising generalizability involves a trade-off between extended performance tasks and briefer structured exercises. That is, breadth of coverage is improved by developing tests that represent a mix of efficient structured items and time-intensive open-ended tasks. Although this may smack of anathema in the context of authentic assessment, it does not mean simply adding standard multiple-choice items to flesh out those aspects of domain knowledge and skill that such a format can reasonably well assess. To begin with, it must be recognized that the contrast between multiple-choice items and open-ended performance tasks is not a dichotomy but, rather, a continuum representing different degrees of response structure. This continuum is variously described as ranging from multiple-choice to student-constructed products or performances (Bennett, Ward, Rock, & LaHart, 1990), for example, or from

multiple-choice to demonstrations and portfolios (Snow, 1992). The main variant is the amount of structure or constraint versus the degree of openness afforded in the response.

There is a wide array of structured item formats toward the multiple-choice end of the continuum. For example, Wesman (1971) describes three varieties of the short-answer form, five varieties of the alternate-choice form, two of the matching form, and eight of multiple-choice, including those allowing more than one right answer. In addition, he discusses three types of context-dependent item sets (the pictorial form, the interlinear form, and the interpretive exercise), to which a fourth type (the problem-solving scenario) has been added (Haladyna, 1992). Thus, contingent sets of structured items can be developed to tap complex aspects of task functioning, such as problem-solving processes and strategies (Ebel, 1984) as well as stylistic learning preferences (Heath, 1964). It should be noted that, contrary to popular misconceptions, structured item formats are not limited to the measurement of fact retrieval. They are also used effectively to assess knowledge application, evaluation skills, and problem-solving proficiencies. Multiple- or forced-choice techniques have also been applied in the measurement of social attitudes, personal needs and motives, vocational interests, aesthetic preferences, and human values (Messick, 1979).

In addition, there are a number of formats at intermediate levels of the continuum, one example being multiple-choice items that require the respondent to give reasons why the chosen option is correct and possibly why each of the unchosen options is incorrect. Another instance is a multiple-rating format in which each of several options is judged for quality against complex standards (Scriven, 1990). Specifically, the student might be asked to read a passage for main idea and then to rate each of four sentences -- say, by marking boxes labeled A to F -- for the quality and completeness with which each captures the main idea. An added requirement might be that if none of the statements receives a grade of B or better, the respondent should write an A-quality main idea sentence of his or her own.

It should be noted that this continuum refers to *response*-form, representing various degrees of structure or constraint imposed on the student's responses. There is another, at least partly independent, continuum

referring to ***stimulus***-form that represents various degrees of structure in the questions or problems presented. These two continua are clearly separable in the structured-stimulus direction because highly structured problems can be presented in either multiple-choice or open-ended formats. The question is the degree to which the two continua are also separable in the unstructured-stimulus direction. In this regard, we should explore the possibility of retaining the efficiency of structured or partly structured responses while simultaneously relaxing the degree of structure in the problems posed. As an instance, patient-management problems might be presented with multiple-choice or key-list options at each decision point. The intent would be to create more realistic, less well-structured problems -- perhaps even ill-structured problems -- having structured or semi-structured response formats. In sum, the aim is to improve domain coverage and generalizability by combining performance tasks with variously structured exercises that tap knowledge and skill relevant to the performance domain.

In the context of performance assessment, the use of structured or semi-structured exercises should attempt to take into account the important specialized validity criterion that J. R. Frederiksen and Collins (1989) refer to as ***transparency*** and Linn and his colleagues (1991) call ***meaningfulness***. Indeed, where appropriate, this specialized validity standard should be applied more generally in educational measurement. The concern is that if the assessment itself is to be a worthwhile educational experience serving to motivate and direct learning, then the problems and tasks posed should be meaningful to the students. That is, students should know what is being assessed and by what methods, as well as why the assessment is worthy of time and effort. Furthermore, the criteria and standards of what constitutes good performance should be clear in terms of both how the performance is to be scored, thereby facilitating student self-assessment, and what steps might be taken or directions moved in to improve performance.

Although meaningfulness and transparency of performance tasks should not be taken for granted, in this regard a number of things can be done (and in many instances are routinely done) to improve somewhat the transparency and meaningfulness of structured and semi-structured exercises. To begin with, structured exercises could be presented in a more contextualized form to make

them more engaging to students. Furthermore, although the actual test exercises are not released in advance for the students to practice, what amounts to test specifications for domain coverage are often provided, along with sample items of each format appearing on the test and recommended strategies for coping with these item-types. Thus, students are told what the test covers and the forms that problems or tasks will take. In addition, they are sometimes provided with complete sample tests and scoring instructions so that students can assess themselves in advance. Moreover, for well-ordered domains or subdomains, scores can be reported in terms of proficiency scales that relate the student's performance level to the kinds of exercises and task processes he or she performs well and to the next more difficult or complex task processes to be mastered.

However, more work needs to be done in communicating to students why this knowledge and skill, assessed in these structured or even unstructured ways, is important and worthy of time and effort. We need to communicate the value of certified competence relevant to the various objectives of schooling as well as the instrumental role of assessed competence in educational decisions affecting the students' future learning and opportunities. This latter reference to educational decisions serves to remind us of the general validity criteria related to substantive and external evidence, namely, that our assessments should not only validly tap the appropriate competency constructs, to be sure, but also have demonstrated utility for the decision-making uses to which they are put.

In regard to the use of structured or semi-structured exercises in conjunction with performance tasks, J. R. Frederiksen and Collins (1989) express the view that subjectivity of scoring, in and of itself, may contribute to the so-called systemic validity of the test. That is, subjectively scored tests may "directly reflect and support the development of the aptitudes and traits they are supposed to measure" (p. 28). This concept of systemic validity is a highly specialized instance of the general validity criterion of social consequences, which holds that the intended and unintended consequences of test use should be consonant with the general testing purposes and that any adverse consequences should not stem from sources of test invalidity such as construct underrepresentation or construct-irrelevant

variance (Messick, 1989). The notion of systemically valid tests is specialized because it focusses on one set of testing consequences among many (Messick 1989), namely, on whether or not the tests "induce curricular and instructional changes in educational systems (and learning strategy changes in students) that foster the development of the cognitive traits that the tests are designed to measure" (J. R. Frederiksen & Collins, 198 , p. 27). Furthermore, interpretation of such teaching and learning consequences as reflective of test validity (or invalidity) assumes that all other aspects of the educational system are working well or are controlled. Thus, in practice the issue may not be just the systemic validity of the tests but, rather, the validity of the system as a whole for improving teaching and learning.

To speculate on the implications of this view that subjective scoring fosters systemic validity, consider that subjectively scored tests might facilitate the development of the skills they are designed to assess in at least two ways. First, tests are usually subjectively scored because they elicit open-ended responses permitting the student to express and exercise complex skills under conditions where the self-feedback of the assessment experience itself, combined with the formal feedback of scores, provides timely information about the current status of task performance and needed improvement. Second, the process of subjective scoring requires reflective analysis and informed judgment which, if carried out by the teacher or in self-assessment by the student, might clarify in detail the standards of good performance. Nuances in the standards might then be more readily related to particular types and directions of improvement to be fostered by instruction or practice. But all this depends on the validity of the feedback and of the educational actions based upon it, which require the consequential evidence of improved learning and performance.

Although these potential benefits of subjective scoring are traded off against the reliability and efficiency of objective scoring, it should be noted that subjectivity is not eliminated in the use of structured exercises. Rather, reflective analysis and professional judgment are shifted from the act of scoring to the act of constructing items and determining answer keys. This suggests that at least some of the consciousness-raising benefits of subjective scoring might be recaptured by having teachers, and possibly

students as well, participate in the development of structured items and their answer keys. More practically, they might engage in reflective evaluation of samples of structured items and their keys relative to specifications for domain coverage. The aim would be to sensitize teachers and students to the nature of the knowledge and skills being assessed, to the standards for what constitutes acceptable answers, and to domain-relevant distinctions between a preferred answer and plausible alternatives. In all of this, it should be remembered that we are not dealing with ideal assessment options, we are dealing with trade-offs and trying to minimize the downside of the trade-off in both directions.

Potential Consequences of Focussing on Tasks versus Constructs

In the discussion thus far, some of the points were made in the language of tasks and behavioral performance, while others were in the language of constructs such as competencies of knowledge and skill underlying the performance. The distinction between competence and performance is an old one, especially in linguistics (Chomsky, 1957). The major point is that although competence must be inferred from observations of performances or behaviors (or from their outcomes or products), this inference is not often straightforward, particularly inferences about lack of competence from poor performance. Indeed, this is the core problem of construct validity, namely, how to establish, via a theoretical integration of convergent and discriminant evidence, that an observed behavioral consistency (as well as relationships of that consistency to other variables) can be accounted for by a particular construct interpretation rather than by plausible rival interpretations.

In education, whether as an objective of instruction or as a target of assessment, we are rarely concerned just with the particular performance per se but, rather, with the knowledge, skill, and other attributes that enable not only the given performance but also a range of other performances engaging the same knowledge and skills. This suggests that constructs like relevant knowledge and skill, rather than domain-relevant tasks and performances ought to drive the development, scoring, and interpretation of performance assessments. Yet, there are arguments on both sides, so we need to ponder

the potential consequences of adopting a task-centered as opposed to a construct-centered approach in performance assessment.

A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. Focussing on constructs also alerts one to the possibility of construct-irrelevant variance which might distort either the task performance or its scoring, or both. With respect to task performance, some aspects of the task may require skills or other attributes having nothing to do with the focal constructs in question, so that deficiencies in the construct-irrelevant skills might prevent some students from demonstrating the focal competencies.

With respect to scoring, construct-irrelevant variance can distort subjective judgments of performance, as when scores on essay tasks focussing on the persuasiveness of arguments, say, or knowledge of biology concepts are influenced in the first instance by quality of handwriting and in the second instance by English-composition skills. What constitutes construct-irrelevant variance, of course, is a central and contentious issue, as witness the complexity of deciding whether English-composition skills would be relevant or irrelevant in judging the persuasiveness of an essay. As a matter of course and especially when in doubt, both the construct-relevant and potentially construct-irrelevant aspects of task performance should be assessed separately so that their differential effects can be taken into account in scoring (Breland, 1991). This broaches the general validity criterion of structural fidelity, which holds that the structure of the scoring model for combining aspects of task performance into construct-relevant scores should parallel the domain structure underlying construct-relevant effects (Loevinger, 1957; Messick, 1989).

Furthermore, the meaning and properties of construct-based scores may be more generalizable across variations in tasks, settings, and population groups

than is task performance per se because of the effort to reduce score contamination by construct-irrelevant variance, which is likely to be task-specific and less generalizable. In any event, generalizability evidence is inherently important in the construct-centered approach because it bears directly on the boundaries of score meaning in relation to the construct meaning that initiates and guides the whole enterprise.

However, a construct-centered approach to performance assessment may be inhibiting in many instances when the constructs at issue are generic proficiencies more akin to construct domains, such as writing, listening, artistic creation, and problem solving. In such instances, constructs of knowledge and skill that contribute to the generic proficiencies do not typically guide the development of scoring criteria and rubrics because they are usually not articulated in advance but, rather, are delineated as part of the process of developing scoring criteria and rubrics. This leads to a task-centered approach to performance assessment which appears to be particularly congenial to fields where the mode of teaching emphasizes repeated demonstration, practice, and critique. In this task-centered approach, as Wiggins (1989) put it, "we must first decide what are the actual performances that we want students to be good at. We must design those performances first and worry about a fair and thorough method of grading them later" (p. 705).

By focussing on an important type of performance and the task that elicits it, one might naturally come to downplay the importance of replicability and generalizability. However, though common, this is by no means a necessary consequence if the issue is kept salient, for example, as Wiggins does in his already mentioned insistence that "multiple and varied tasks are required" to sustain interpretations of competence. Nor does the construct-centered approach guarantee attention to generalizability. For example, the testing system favored by J. R. Frederiksen and Collins (1989) begins by specifying the construct to be assessed, which is linked both to the goals of teaching and learning and to a primary-trait approach to scoring. But they then invoke an Olympic Games assessment metaphor which tacitly stresses the performance per se as the target of assessment and, as has been seen, generalizability is simply not raised as an issue.

Furthermore, in the task-centered approach, by virtue of focussing on tasks, the concept of irrelevant variance in task performance loses all meaning because every skill required, however task-specific, is relevant to task completion. Nonetheless, the concept of construct-irrelevant variance still has force with respect to test scoring because, once multiple constructs such as clarity of expression or depth of understanding are delineated, judgments of one construct can contaminate judgments of the other. More subtly, observed though unscored aspects of task performance can contaminate judgments of scored aspects.

By delineating de novo for each task not only the aspects of performance quality to be scored but also the contributing constructs of knowledge and skill, the task-centered approach is in danger of tailoring scoring criteria and rubrics to properties of the task and of representing the constructs in task-dependent ways that limit generalizability. This is in contrast to tailoring scoring criteria and rubrics to properties of the constructs as they might be revealed across varied tasks. Task-centered scores could lead to a proliferation of task-dependent constructs in much the same way that operational definitions led to a proliferation of constructs tied to specific measurement operations. Thus, the preemptive emphasis on tasks and performances in the task-centered approach to performance assessment may not only bring behaviorism back into education by the rear door but, in effect, also behaviorism's talisman and shield, the operational definition.

On the other hand, construct-centered scoring criteria and rubrics are prey to the opposite danger of becoming too generic. This is especially problematic if the construct in question can legitimately have multiple manifestations at different levels of performance quality. The scoring rubrics need to be richly detailed enough to cope flexibly with this multiplicity without being task specific. One possibility is to aim for scoring rubrics that are neither specific to the task nor generic to the construct but, rather, are in some middle ground reflective of the classes of tasks that the construct empirically generalizes or transfers to. One benefit of such midlevel scoring rubrics is that they may provide a basis for feedback that is likely to be more informative than a "C" in American History, say, and more generally useful than a graded essay on economic causes of the Civil War.

In any event, one of the advantages of the construct-centered approach is that the focal constructs can help guide the selection or construction of tasks that would optimally reveal them. This might prove especially useful in constructing so-called authentic assessments which, according to Wiggins (1989), "replicate the challenges and standards of performance that typically face writers, businesspeople, scientists, community leaders, designers, or historians" (p. 703-704). That is, authentic tasks are "representative of the ways in which knowledge and skills are used in 'real world' contexts" (J. R. Frederiksen & Collins, 1989, p. 20). Shepard (see Kirst, 1991) goes even further by indicating that authentic "assessment tasks themselves are real instances of extended criterion performances" (p. 21).

The issue here appears to be the extent to which the criterion situation is faithfully simulated by the test. In the task-centered approach to authentic assessment, credibility depends on the simulation of as much real-world complexity as can be provided. But along with realism there comes a multiplicity of variables as well as lack of control. This puts an enormous burden on test development: It makes difficult the creation and application of criteria for scoring relevant aspects of this complexity, while at the same time jeopardizing scorer reliability and very likely limiting generalizability. The construct-centered approach makes possible a compromise by focussing on selected constructs of knowledge and skill and the conditions of their realistic engagement in task performance.

For a given cost and level of control, there are trade-offs in simulation between ***comprehensiveness*** (or the range of different situational aspects that are simulated) and ***fidelity*** (or the degree to which each simulated aspect approximates its real-world counterpart). What is important to simulate are the critical aspects of the criterion situation that elicit those performances from which the focal constructs of knowledge and skill are inferred, at a sufficient level of fidelity to detect relevant differences and changes in the focal performance variables (Fitzpatrick & Morrison, 1971). Other aspects of the test situation can be controlled or standardized. Such simulated tasks are authentic in that they replicate the challenges and standards of real-world performances and are representative of the ways in which knowledge and

skills are used in real-world contexts, even though they do not simulate all of the complexity of real-world functioning.

The term "representative" has two distinct meanings here, both of which are applicable. One is in the cognitive psychologist's sense of representation or modeling, the other is in the Brunswikian sense of ecological sampling (Brunswik, 1956; Snow, 1974). The choice of tasks and contexts to simulate is a sampling issue. The comprehensiveness and fidelity of simulating the construct's realistic engagement in performance is a representation issue. Both issues are important in authentic assessment. Indeed, they mirror the concerns of the content, substantive, and generalizability aspects of construct validity.

AUTHENTICITY AND DIRECTNESS AS VALIDITY STANDARDS

The notion of authenticity in testing is a many-faceted concept. In general, authentic assessments involve the creation of products and "the performance of tasks that are valued in their own right" (Linn et al., 1991, p. 15), which is not quite the same as task performance that reveals valued knowledge and skills or other constructs. One implication is that the knowledge and skills are valued only because they are instrumental to the performance of an important task. However, from another perspective, the knowledge and skills are valued because they are ***enabling*** variables instrumental to the performance of a variety of tasks or to the creation of a variety of products. This confronts us once again with the task-centered as opposed to the construct-centered view of performance assessment. In any event, exposure to authentic assessment is expected to provide the student with a meaningful educational experience that facilitates learning and skill development as well as deeper understanding of the requirements and standards for good performance.

Specifically, authentic assessments aim to capture a richer array of student knowledge and skill than is possible with multiple-choice tests; to depict the processes and strategies by which students produce their work; to align the assessment more directly with the ultimate goals of schooling; and, to provide realistic contexts for the production of student work by having the

tasks and processes, as well as time and resources, parallel those in the real world (Arter & Spandel, 1992). Moreover, for instructional assessment, some additional aims are to provide ongoing student feedback encouraging efforts toward growth and promoting active student engagement in both learning and the control of learning, thereby integrating assessment with instruction.

Yet, in spite of these clear intentions, a fundamental ambiguity pervades authentic educational assessments, namely, authentic to what? If the content of a portfolio mirrors the emphasis in the curriculum and classroom, is that *in*authentic (Arter & Spandel, 1992)? Should educational assessments be authentic reflections of classroom work or authentic representations of real-world work? Or should the issue be finessed by insisting that classroom work should authentically reflect real-world activities?

The expressed aims of authentic assessments point to a number of intended consequences that need to be evaluated, along with unintended consequences, to appraise the validity of authenticity as a measurement construct. Before considering the consequential basis for the validity of performance assessment, however, let us first probe the concept of authenticity more deeply to see what it might mean in measurement terms and what kinds of evidence, in addition to evidence of consequences, might sustain authenticity as a validity standard. In so doing, we emphasize two properties: first, task contextualization, as opposed to the decontextualized assessment attributed to multiple-choice items; second, complexity of task functioning, as opposed to the separate assessment of decomposed skills presumed to be required in structured exercises. In addition, the concept of directness of assessment, which typically goes hand in hand with authenticity, is also probed to see what it might mean in measurement terms and what kinds of evidence might sustain directness as a distinct validity standard.

Contextualized versus Decontextualized versus Cross-Contextual Assessment

According to Resnick and Resnick (1991), decontextualized assessments assume that "each component of a complex skill is fixed, and that it will take the same form no matter where it is used" (p. 43). This statement makes reference to both a complex skill and its components, so it may lead to some confusion between decontextualization and decomposability. To avoid this

confusion, we are dealing in this section only with one or the other, only with complex skills or only with component skills, but not both at the same time. To be sure, the behavior from which the skill is inferred is a function of the task and the context as well as the person. Hence, the form in which the skill is revealed in behavior is subject to interactions with task and context variables.

Consider Cronbach's (1989) automotive example:

> The variables engineers use to describe automobile performance are functionally related to the octane rating of the fuel. What the functions will be depends upon the engine design, the cleanness of the engine, and the driving speed. These complications are matters for the engineer to understand, but the variation of the parameters does not per se call the validity of octane measurement into question. (p. 158)

So it is with measures of cognitive skill. Thus, contrary to Resnick and Resnick, there is no necessary assumption that the skill takes the same form in all contexts. What is important is not that the skill appears different in different contexts, but that it changes nonramdomly with conditions and hence correlates with construct-relevant variables.

Although interactions with context are inevitable, there are a number of ways of coping with them in measurement. One approach attempts to strip the problem context of all irrelevancies, retaining only the task information needed to engage the focal knowledge and skills involved in task processing. Such attempts at decontextualization rarely try to be absolute and often include information to pinpoint particular difficulties in task solution, as in providing plausible though flawed alternatives to the preferred answer, whether in the form of extraneous material in a reading passage or of distractors in a multiple-choice item. Another approach attempts to draw inferences about skill not from behavior in context, but from consistencies in behavior across contexts or across varied tasks within context. On a small scale, this is what test specifications do when they vary item-types or the knowledge domains of reading passages or the genres of writing samples. It is akin to what Humphreys (1962) called the achievement of score homogeneity by the control of heterogeneity. A third approach tries to model the skill as a

function of parametric variations in dimensions of task difficulty and contextual influence. As a final instance, one could treat the skill as revealed in different contexts as qualitatively different skills, which is where Resnick and Resnick appear to come down.

They hold that "knowledge and skill cannot be detached from their contexts of practice and use. . . . That means, in turn, that we cannot validly assess a competence in a context very different from the context in which it is practiced and used" (Resnick & Resnick, 1991, p. 43). In effect, this appears to bring us once again to a behavioristic proliferation of skill constructs that are qualitatively different in different contexts of practice and use. This results in as many distinct skills operationally as there are skill-context combinations. The alternative to this neo-behaviorist view is not necessarily decontextualization but, rather, cross-contextual measurement. In this regard, the finding of modest correlations across task types or contexts -- for example, as between writing a persuasive letter and writing a procedural explanation -- should not too quickly be taken to mean a lack of generalizability across genres. Rather, it indicates how large a sample of varied tasks or contexts is needed to generalize with any confidence across genres or to a broader construct domain. Thus, the important contrast may not be between contextualized versus decontextualized assessment but between contextualized versus cross-contextual assessment.

These two contrasts may represent two distinct continua or perhaps a single complex continuum. For example, one of the goals of instruction associated with the use of models in mathematics and science has been described in terms of development "from situated knowledge to decontextualized understandings" (Lesh & Lamon, 1992, p. 7). This suggests that cross-contextual generalizability may fall in an intermediate range between the extremes of context-dependence and decontextualization or abstractness -- in cognitive processing as well as in measurement. In this connection, the domain-specific knowledge-and-skill structures or schemas characteristic of expert performance do not necessarily contradict such a complex continuum because they are tied not to problem contexts within domain but to perceived deep structures cutting across surface contexts (Chi, Glaser, & Farr, 1987; Dreyfus & Dreyfus, 1986).

This possibility of a developmental continuum, highly speculative though it may be, is broached here merely to signal potential contingencies in the applications of measurement. That is, whether one focusses on context-dependent measurement (or instruction) or on increasingly context-generalizable measurement or on abstract or context-independent measurement may depend on the student's level of developing expertise in the subject matter. In any event, one important reason for favoring rich contextualization of problems or tasks is to engage student interest and thereby improve motivation and effort. The cross-contextual approach does not dispute this, but merely requires multiple and varied contextualized tasks.

Nevertheless, we should not take it for granted that richly contextualized assessment tasks are uniformly good for all students. There are very few one-edged swords in the measurement enterprise, and contextualization is unlikely to be one of them. Indeed, contextual features that engage and motivate one student and facilitate effective task functioning may alienate and confuse another student and bias or distort task functioning. Given consistent findings from analyses of differential item functioning that certain types of content or subject matter in reading passages or in algebra word problems or other item types have different performance consequences as a function of gender and ethnicity (O'Neill & McPeek, 1992; Schmitt & Dorans, 1990), it would not be surprising if other contextual features had similar effects. In any event, such potential unintended consequences would need to be evaluated as part of the consequential basis for the validity of contextualized assessment. Since this cannot be undertaken without multiple and varied contextualized measures, we accumulate one more argument in favor of cross-contextual assessment.

One approach to coping equitably with differential student responsiveness to context is to develop an aggregate measure of the construct across a variety of item contexts in an effort to balance the effects of different student backgrounds and interests. This is in the spirit of cross-contextual assessment. Another approach is to develop multiple test forms, each one assessing the construct in a different context, such as tests of comprehension or reasoning in the biological sciences, the physical sciences, the social sciences, or the humanities. The appropriate form of the construct-in-context

could then be matched to the student's background, perhaps allowing student choice as to which context would best reveal his or her reasoning skills.

Indeed, in many applied instances such as licensure or the prediction of job performance, the immediate concern is not with generalizable measures of the focal construct but with measures of the construct-in-context, for example, as captured in work samples or simulations. Generalizable construct measures become important, of course, if the criterion domain is extensive or if the job or conditions are likely to change. The issue of contextualization thus becomes increasingly more complex. The continuum, if such it be, now appears to range from context-dependent measurement, to the assessment of constructs-in-context, to cross-contextual assessment, to decontextualized or abstract assessment.

However, there is another perspective on contextualized versus decontextualized assessment that is related to our earlier question of "Authentic to what?" For example, Sternberg (1990) maintains that it is not that standardized tests are decontextualized but, rather, that "the items are contextualized with respect to the environment in which they are supposed to measure performance -- namely, the environment of the school" (p. 211). Jackson (1968) reminds us that students live a real *Life in Classrooms* and Sternberg (1991) characterizes the cognitive demands of that life as follows:

> Most of the problems students need to solve in school are relatively short, contain relatively little information about the problem situation, have a single correct answer, are disconnected from each other, contain little or no 'real world' content, are prestructured, are solved individually, and are well-structured. Therefore, . . . the issue is that test problems are contextualized with respect to school problems, but that both test problems and school problems are decontextualized with respect to every day life problems. (p. 212)

As a consequence, Sternberg (1990) calls for cross-contextual testing using items that tap both academic and practical contexts, but the practical items in his *Triarchic Abilities Test* are still discrete, structured, and multiple-choice in format.

Given the dilemma of being authentic either to the school or to the real world, there is of course the alternative of making school activities more authentic to real-world activities and using authentic educational assessment as a vehicle for achieving this. To be realistic, however, it is unlikely that such a transformation could be accomplished completely not just because of logistical problems and inadequate resources, but because the cognitive demands of schooling described by Sternberg also serve legitimate pedagogical purposes. Such discrete instructional aims also have a long history in education predating the introduction of multiple-choice technology. One possibility, for example, is that some so-called decontextualized approaches in instruction and assessment may facilitate the development of abstraction and inference skills, which in turn may facilitate the development of domain-specific knowledge-and-skill structures as well as abstract thinking skills characteristic of expert performance. According to Gardner (1983),

> one learns in school to deal with information outside of the context in which it is generally encountered; to entertain abstract positions and to explore the relations among them on a hypothetical basis; to make sense of a set of ideas, independent of who says them or of the tone of voice in which they are said; to criticize, to detect contradictions, and to try to resolve them. One also acquires a respect for the accumulation of knowledge, for ways to test statements in which one does not have an immediate interest, and for the relationship between bodies of knowledge that might otherwise seem remote from one another. This valuing of abstract concerns, which relate to reality only by a lengthy chain of inference, and a growing familiarity with "objective" writing, reading, and testing eventually spawns a person at home with the principles of science and mathematics and concerned about the extent to which his views and behavior accord with these somewhat esoteric standards. (p. 164)

This brings us back to a strategy of assessment that combines complex performance tasks toward the contextualized end of the continuum with structured exercises toward the decontextualized end, but with both now being viewed as authentic to a working mix of real-world and school activities.

It should be clear from all of this that contextualization is too complexly intertwined with the purposes of assessment and perhaps with student

levels of developing expertise to serve as an unequivocal touchstone for authenticity as a validity standard. We turn next to the property of skill complexity versus decomposition of skilled performance to see if it might serve as a more credible touchstone.

Assessment of Complex versus Decomposed Skills

According to Resnick and Resnick (1991), assessments that decompose complex competencies or generic proficiencies into their component skills for separate measurement fail to recognize that "complicated skills and competencies owe their complexity not just to the number of components they engage but also to interactions among the components and heuristics for calling upon them" (p. 42). This is clearly a fundamental point which is rarely if ever in dispute. What is in dispute are the action implications that are drawn from it.

For instance, does it necessarily follow that "efforts to assess thinking and problem-solving abilities by identifying separate components of those abilities and testing them independently will interfere with effectively teaching such abilities" (Resnick & Resnick, 1991, p. 43)? Does this mean that measurement should shift completely away from separate assessment of the component skills to sole assessment of the complex skill as a functioning whole? What about the alternative view that "subprocesses need to be assessed independently so that test takers will direct their efforts to doing well in all phases of the task domain being tested" (J. R. Frederiksen & Collins, 1989, p. 30)? What about assessing both the complex skill and its component skills, including metacognitive components for organizing and regulating the component skills and for planning, monitoring, and evaluating the complex performance (Sternberg, 1985)? Does it necessarily follow that assessing, teaching, and practicing component skills in isolation or in various combinations is deleterious, even in connection with the diagnosis and remediation of complex-skill difficulties? Might not assessment of component skills help one to understand the nature of the complex skill and the sources of its complexity, providing a functional basis for improving methods of teaching?

Focussing on component skills may indeed be deleterious if it means that effective teaching and practice of the complex skill is forgone, but why not do both as in the teaching of instrumental music or sports or even thinking skills? In connection with thinking skills, for example, Sternberg's (1985) separate appraisal of the components of reasoning as well as those of verbal comprehension led him to emphasize these cognitive and metacognitive components in the training of thinking. But he also combined these components to model global task performance and, though the models fit reasonably well, there was typically task variance left unaccounted for. This might mean that more component skills need to be identified, but it might also indicate that complex interactions among the components and the metacognitive heuristics contribute to facile overall skill functioning over and above the sum of the components. The important point in all this is not that the complex skill is decomposed into components for separate measurement but that, in the process, something is left out. That is, the complex-skill construct is underrepresented in componential measurement. Furthermore, the action implication of this underrepresentation that is favored by most advocates of authentic testing is that performance assessment of the complex skill as a functioning whole guarantees that nothing important will be left out.

Concern about leaving things out in the teaching and assessment of complex knowledge and skills is at the heart of the authentic testing movement. Things that are left out are perceived to be unvalued, so they are likely to be unattended to and hence underdeveloped. This perceived devaluing of complex skills in favor of component skills in educational testing, and ultimately in teaching, is what energizes the movement. But in moving to redress this perceived devaluation, is it better to swing completely over to the performance assessment of overall skills or to attempt to measure the complex functioning, to be sure, but the component skills as well?

Much depends on whether or not the performance assessment of complex-skill functioning is indeed a guarantee that no important aspect of the skill construct is left out. No test can ever completely capture the construct, of course, because the construct refers not just to specific tasks but to processes and other attributes underlying a domain of potential tasks. But even though we can never capture the construct completely in any measurement

instance, we should strive not to leave any important aspect of the construct out. This corresponds to the general validity standard of minimizing construct underrepresentation, which requires convergent and discriminant evidence bearing on any and all aspects of construct validity, namely, the content, substantive, structural, external, generalizability, and consequential aspects mentioned earlier (Messick, 1989).

The problem is that one may be reasonably confident that nothing is left out when inferring the complex skill from well-crafted products or from outstanding performances. But what about inferences from intermediate level or only moderately good performances or especially from poor performances? Is the performance poor because the component skills were not facilely deployed, or because one or more of the component skills is deficient, or because of inattention or low motivation? Does the performance assessment tacitly assume that the component skills have achieved necessary minimum levels? Answers to these questions should radically influence strategies for student improvement. But if we cannot answer these questions, how can we be assured that no important part of the skill construct is left out?

To complicate matters further, what is critical in performance assessment, as in all assessment, is not what is operative in the task performance but what is captured in the test score and its interpretation. By what evidence can we be assured that the scoring criteria and rubrics used in holistic, primary trait, or analytic scoring of products or performances capture the fully functioning complex skill? Or should construct interpretations be explicitly limited to those aspects of skill functioning that the scoring criteria cover? Wiggins (1989) indicates that "authentic tests use multifaceted scoring systems instead of a single aggregate grade. The many variables of complex performance are disaggregated in judging" (p. 711). But component skills are difficult to disentangle from complex functioning by judgment alone, and the multifaceted scores often reflect aspects of performance quality, such as clarity of expression or degree of coherence, rather than aspects of skill. Indeed, the complexities of "realistic" performance tasks may make the component processes more not less opaque. Hence, it is by no means clear that the whole complex skill is

automatically captured in performance assessment. In any event, convergent and discriminant evidence is needed to sustain complex score interpretation.

In the performance assessment of complex skills as ordinarily conducted, we once again stumble over the legacy of behaviorism. As Cronbach (1989) put it,

> operationalism lingers on in the hearts of some specialists in achievement testing. . . . For them, interpretation begins and ends with a highly specific definition of the kinds of tasks the examinee should be faced with; a test sampling from the defined domain is valid by fiat. This program is coherent but shortsighted. For understanding poor performance, for remedial purposes, for improving teaching methods, and for carving out more functional domains, process constructs are needed (p. 161).

Hence, process constructs need to be assessed -- not instead of but, rather, in addition to complex performances.

The basic point in this discussion of complex and component skills is that the validity standard implicit in the concept of authenticity appears to be the familiar one of construct representation (Embretson, 1983; Messick, 1989). That is, evidence should be sought that the presumed sources of task complexity are indeed reflected in task performance and that the complex skill is captured in the test scores with minimal construct underrepresentation. For example, in measuring analytical skill, it matters whether the generation of a mathematical proof reflects analytical thinking or the reproduction of a memorized sequence of steps. It also matters whether the test scores, especially intermediate and low scores, reflect analytical skill or mathematical knowledge, or a complex of both. Among the validity criteria specialized for performance assessment that Linn and his colleagues (1991) proposed is a closely related one which they call ***cognitive complexity***. This criterion entails "an analysis of the cognitive complexity of the tasks and the nature of the responses that they engender" (p. 19). However, it is not that the cognitive complexity of performance tasks should necessarily be higher than that of structured tasks, because the complexity of the latter can be very high indeed. It is that the level and sources of task complexity should match those of the construct being measured and be attuned to the level

of developing expertise of the students assessed. Let us now turn to the handmaiden of authenticity, namely, directness of assessment, to see what validity standard, if any, is implicit in it.

Direct Assessment of Task Behaviors, Indirect Assessment of Skills

J. R. Frederiksen and Collins (1989) invoke ***directness*** as a specialized validity standard for performance assessment and, as is common in the field, they speak of the direct assessment of knowledge and skills. However, constructs of knowledge and skill cannot be assessed directly but, rather, are inferred from performances and products. Technically, even attributes of task behavior and aspects of performance quality are not typically assessed directly because the performance scores are mediated by judgment, whether human or artificial. Indeed, the term "direct assessment" is generally inappropriate, especially in the behavioral sciences, and should not be used. It always claims too much: It runs afoul of Heisenbergian uncertainties in measuring dynamic systems and ignores intervening processes in measuring static systems. Nonetheless, the term is being widely used in some educational circles, so that one must now ask what might be meant by direct assessment?

One possibility is that direct assessment refers to task conditions in which the student can freely express the complex skill at issue unfettered by structured item forms or restrictive response formats. This implies that direct assessment is assessment that employs open-ended tasks and judgmental scoring, again by means of either humans or automated algorithms. The intent of such testing is to minimize constraints on student behavior associated with sources of construct-irrelevant method variance such as testwiseness in coping with various item-types, tendencies toward guessing, and other artificial restrictions on students' representations of problems or on their modes of thinking or response. In the case of authenticity, the concern was with not leaving anything relevant out, leading to construct representation or minimal construct underrepresentation as the implicit validity standard. In the case of directness, the concern appears to be with not putting anything irrelevant in, leading to minimal construct-irrelevant variance as the implicit validity standard.

However, in direct assessment there is a trade-off between method variance that affects student performance and method variance that affects the subjective scoring of that performance. Hence, we need careful training, calibration, and monitoring of observers or judges not just to achieve adequate scorer reliability, but also to attenuate the effects of a wide array of observation and judgment biases and other sources of method variance in scoring (Cattell, 1968, 1977; Messick, 1983; Webb, Campbell, Schwartz, & Sechrest, 1966; Weick, 1968). We also need to take account of observer-interference effects whereby the students' spontaneous behaviors may be altered as a consequence of their awareness of being observed and of their attendant interactions with characteristics of the observer and the mode of observation (Weick, 1968). This latter type of method variance, unfortunately, transcends the trade-off and potentially distorts both the observer's scoring judgments and the student's task behavior.

Nor should it be taken for granted that open-ended tasks are unstructured or devoid of the constraints of method variance. For example, in an exemplary authentic test described by Wiggins (1989), an oral history on a topic of the student's choice was to be completed based on interviews with four appropriate people as sources. The student was to create three workable hypotheses based on preliminary investigations and formulate four questions to be asked to test each hypothesis. This is hardly an unstructured task nor is it likely to be devoid of method variance that constrains at least some students' modes of thinking and response. But to attempt to minimize construct-irrelevant method variance one needs to know what focal constructs are being assessed, so that either controls for method variance can be introduced or method effects can be taken into account in score interpretation.

With respect to validity standards implicit in authenticity and directness as measurement concepts, then, we find that the former requires evidence that the test is not unduly narrow because of missing construct variance, whereas the latter requires evidence that the test is not unduly broad because of added method variance. We thus return to the need for convergent and discriminant evidence to counter the two major threats to construct validity, namely, construct underrepresentation and construct-irrelevant variance. If authenticity and directness, respectively, serve as

popular watchwords for countering these two threats, this is all to the good, so long as the watchwords are signals for forthcoming evidence.

Nor is this evidential basis all that is needed to move performance assessment vigorously into established practice, especially given its problems of cost and efficiency. Incidentally, Linn and his colleagues (1991) propose ***cost and efficiency*** as a specialized validity criterion because these issues cannot be ignored in applications of performance assessment. Yet, validity of performance tests should not be conceived in terms of improved costs and efficiency alone but, rather, in terms of costs and efficiency relative to benefits, which is the general external validity criterion of utility (Cronbach & Gleser, 1965; Messick, 1989). In addition to the evidential basis, we also need a consequential basis for the validity of performance tests, especially since they trade so much on their potential consequences for improving teaching and learning. Evidence of such positive consequences in conjunction with little or no adverse consequences would make performance testing the coin of the realm in educational measurement.

Evaluating Both Intended and Unintended Testing Consequences

The consequential basis of test validity includes evidence and rationales for evaluating the intended and unintended consequences of test interpretation and use in both the short- and long-term. Particularly prominent is the evaluation of any adverse consequences for individuals and groups that are associated with bias in test scoring and interpretation or with unfairness in test use. Similar issues are broached by Linn, Baker, and Dunbar (1991) in terms of their specialized validity criteria of ***consequences*** and ***fairness***.

The primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups, especially gender and racial/ethnic groups, should not derive from any source of test invalidity such as construct underrepresentation or construct-irrelevant variance (Messick, 1989). That is, low scores should not occur because the test is missing something relevant to the focal construct which, if present, would have permitted the affected students to reveal their competence. Furthermore, low scores should not occur because the test contains something irrelevant that interferes with the affected students' demonstration of competence. To

the extent that performance tests exhibit less construct underrepresentation and construct-irrelevant method variance, as authenticity and directness promise, then one might expect less adverse impact associated with sources of test invalidity. But this does not mean that there would necessarily be less adverse impact associated with the valid description of existing group differences. One must also be alert to the possibility of unintended consequences of performance assessment such as increased adverse impact for gender and racial/ethnic groups because of short-term misalignments in their educational experiences vis-à-vis the different cognitive demands of authentic testing and teaching. If found, one should monitor the situation to see how short-term it is likely to be and what resources are needed to redress the new imbalance.

In assembling evidence of consequences, it is always the unintended consequences that are hardest to deal with because it is difficult to know where to look for unintended effects. However, just because effects are unintended does not mean that they cannot be anticipated and, if not attenuated by introducing changes or controls in the testing procedures, at least evaluated and taken into account in test interpretation and use. In this regard, the construct-centered approach to performance assessment may have some advantages because the meaning of the construct to be assessed provides a rational basis for hypothesizing potential testing outcomes and for anticipating possible side effects. That is, the construct's theory, by articulating links between processes and outcomes, provides clues to possible effects (Messick, 1989).

Another route to locating unintended effects is to search for potential corruptions of the testing purposes, such as teaching to the ***assessed*** version of the learning goals rather than to the broader constructs reflected in the stated goals or, if testing tasks are prespecified, teaching specifically to the tasks (see Shepard's comments in Kirst, 1991). In the first instance, the curriculum can be corrupted and, in the second, score validity can be corrupted, as when a performance test of analytical thinking comes to reflect memorization of procedures. Another strategy for dealing with unintended effects is to systematically view testing outcomes in terms of multiple value

perspectives, because one person's intended positive effect may be another person's deleterious side effect (Messick, 1989).

OVERVIEW AND ADMONITION

The interplay between evidence and consequences in the validation of performance assessments underlies and sustains an interplay between score meaning and social values in the justification of score interpretation and use. The validity issues, being many-faceted and intertwined, are difficult to disentangle, which is why test validity has come to be viewed as a unified concept. Nonetheless, let us now attempt to reprise some major points, recognizing that their separation from qualifying arguments makes them sound more authoritative than is warranted. In one sense I do this purposely in order to offer a strong position to be challenged as we cope with the problems of validation and use of performance assessments for accountability purposes.

To begin with, the interpretation and use of performance assessments, like all assessments, should be validated in terms of content, substantive, structural, external, generalizability, and consequential aspects of construct validity. These general validity criteria can be specialized for apt application to performance assessments, if need be, but none should be ignored.

Decisions to include particular student performances or products in the assessment should be rationally tied to the purposes of the testing, to the nature of the substantive domain at issue, and to construct theories of pertinent skills and knowledge. Furthermore, one should be clear about whether the performance or product per se is the target of the assessment as opposed to serving as a vehicle for the assessment of knowledge, skills, or other constructs, because the implications for validation are strikingly different.

Given the trade-offs in performance assessment between domain coverage and generalizability on the one hand and time-intensive depth of examination on the other, assessment batteries ought to represent a mix of efficient structured exercises and less structured open-ended tasks. Efforts should also be made to refine task formats at intermediate levels of structure as

well as item forms that relax task-structure while retaining, to the degree possible, the scoring objectivity and efficiencies of response-structure.

Where possible, a construct-driven approach to performance assessment rather than a task-driven approach should be adopted because the meaning of the construct guides the selection or construction of relevant tasks as well as the rational development of scoring criteria and rubrics. Focussing on constructs also makes salient the issues of construct underrepresentation and construct-irrelevant variance, which are the two main threats to validity.

The conflict between contextualization and decontextualization of problems or tasks should be resolved by recognizing that both, in their own ways, can serve legitimate instructional and measurement purposes. The problem is to assure that the forms and purposes of the testing are appropriately matched. In any event, a multiplicity of problem contexts should be employed to facilitate cross-contextual assessment and the appraisal of generalizability or the lack thereof.

For comprehensive assessment, both complex skills and their component skills, where delineated, should be tested. To emphasize one at the expense of the other invites construct underrepresentation as well as difficulties in diagnosis and remediation.

The validity standard implicit in authenticity of assessment as a measurement concept is the familiar one of construct representation or minimal construct underrepresentation. The validity standard implicit in directness of assessment is minimal construct-irrelevant method variance. Together, they signal the need for convergent and discriminant evidence that the test is neither unduly narrow because of missing construct variance nor unduly broad because of added method variance.

Finally, evidence of intended and unintended consequences of test interpretation and use should be evaluated as an integral part of the validation process. This evidence should especially address both the anticipated positive consequences of performance assessment for teaching and learning as well as potential adverse consequences bearing on issues of bias and fairness. In particular, evidence is needed to assure that any adverse testing consequences do not derive from sources of test invalidity.

In closing, we note that a strong case can be made that the creative ferment energizing the performance testing movement was stimulated by Norman Frederiksen's (1984) consciousness-raising article on *The Real Test Bias*. This article traced the influences of testing on teaching and learning, especially the potentially deleterious educational effects of multiple-choice testing. But N. Frederiksen's manifesto is a morality tale, and like all good morality tales the salient moral is not the only or even the most important one. It is not just that some aspects of multiple-choice testing may have adverse consequences for teaching and learning, but that some aspects of all testing, even performance testing, may have adverse educational consequences. And if both positive and negative aspects, whether intended or unintended, are not meaningfully addressed in the validation process, then the concept of validity loses its force as a social value.

REFERENCES

Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice*, 11(1), 36-44.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). Toward a framework for constructed-response items (ETS RR 90-7). Princeton, NJ: Educational Testing Service.

Breland, H. (1991). *A study of gender and performance in advanced placement history examinations* (CB 91-4, ETS RR 91-61). Princeton, NJ: Educational Testing Service.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Cattell, R. B. (1968). Trait-view theory of perturbations in ratings and self ratings (L(BR)- and Q-data): Its application to obtaining pure trait score estimates in questionnaires. *Psychological Review*, 75, 96-113.

Cattell, R. B. (1977). A more sophisticated look at structure: Perturbation, sampling, role, and observer trait-view theories. In R. B. Cattell & R. M. Dreger (Eds.), *Handbook of modern personality theory* (pp. 166-220). New York: Wiley.

Chi, M., Glaser, R., & Farr, M. (1987). *Nature of expertise*. Hillsdale, NJ: Erlbaum.

Chomsky, N. (1957). *Syntactic structures*. The Hague, Netherlands: Mouton.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy -- Proceedings of a symposium in honor of Lloyd G. Humphreys* (pp. 147-171). Chicago: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. New York: Free Press.

Ebel, R. L. (1984). Achievement test items: Current issues. In B. S. Plake (Ed.), *Social and technical issues in testing: Implications for test construction and usage*. (pp. 141-154). Hillsdale, NJ: Erlbaum.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, **93**, 179-197.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 237-270). Washington, DC: American Council on Education.

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

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, **39**, 193-202.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice*, **11**(1), 21-25.

Heath, R. W. (1964). Curriculum, cognition, and educational development. *Educational and Psychological Measurement*, **24**, 239-253.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist*, **17**, 475-483.

Jackson, P. W. (1968). *Life in classrooms*. New York: Holt, Rinehart and Winston.

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

Lesh, R., & Lamon, S. J. (1992). Assessing authentic mathematical performance. In R. Lesh & S. J. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp.). Washington, DC: American Association for the Advancement of Science.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York: Macmillan.

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

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

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

Messick, S. (1964). Personality measurement and college performance. *Proceedings of the 1963 Invitational Conference on Testing Problems* (pp. 110-129). Princeton, NJ: Educational Testing Service. [Reprinted in A. Anastasi (Ed.). (1966). Testing problems in perspective (pp. 557-572). Washington, DC: American Council on Education.]

Messick, S. (1965). Personality measurement and the ethics of assessment. *American Psychologist*, 20, 136-142.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955-966.

Messick, S. (1979). Potential uses of noncognitive measurement in education. *Journal of Educational Psychology*, 71, 281-292.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, *10*(9), 9-20.

Messick, S. (1983). Assessment of children. In P. H. Mussen (Ed.), *Handbook of child psychology* (Vol. I, 4th ed.): W. Kessen (Ed.), *History, theory, and methods* (pp. 475-526). New York: Wiley.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Erlbaum.

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

Messick, S. (1992a). Construction and validation of meaning and implications for action in instructional assessment. In G. A. Forehand, C. K. Tittle, & S. Messick, *Assessment and the construction of meaning for instruction and learning* (ETS RM-92-2). Princeton, NJ: Educational Testing Service.

Messick, S. (1992b). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., pp.). New York: Macmillan.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational Measurement* (pp. 445-474). New York: Macmillan.

O'Neill, K., & McPeek, W. M. (1992). Item characteristics that are associated with differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice (pp.)*. Hillsdale, NJ: Erlbaum.

Resnick, L. B., & Resnick, D. P. (1991). 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. 37-75). Boston: Kluwer.

Schmitt, A., & Dorans, N. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, 27, 66-81.

Scriven, M. (1990 unpublished manuscript). Multiple-rating items.

Snow, R. E. (1974). Representative and quasi-representative designs for research on teaching. *Review of Educational Research*, 44, 265-291.

Snow, R. E. (1992). Construct validity and constructed response tests. In R. E. Bennett & W. C. Ward, Jr. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp.). Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J. (1990). T & T is an explosive combination: Technology and testing. *Educational Psychologist*, 25, 201-222.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: A survey of non-reactive research in social science*. Chicago: Rand McNally.

Weick, K. E. (1968). Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. II, 2nd. ed., pp. 357-451). Reading, MA: Addison-Wesley.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 81-129). Washington, DC: American Council on Education.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 79, 703-713.

Wolf, D., Bixby, J., Glenn, J., III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education*, 17, 31-74.