Performance Puzzles: Issues in Measuring Capabilities and Certifying Accomplishments

CSE Technical Report 415

Lauren Resnick
CRESST/University of Pittsburgh, LRDC

June 1996

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

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

The findings and opinions expressed in this report do not reflect the policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

PERFORMANCE PUZZLES: ISSUES IN MEASURING CAPABILITIES AND CERTIFYING ACCOMPLISHMENTS[1]

Lauren B. Resnick

CRESST/University of Pittsburgh, LRDC

Abstract

This article explores issues involved in using assessments as a means of defining standards and encouraging efforts to meet them. It compares the European examination system with the present American testing system. It also considers issues that must be faced in defining learning domains in ways that do not encourage narrowly focused training on specific assessment items.

Performance assessment is on the rise. What was, just a few years ago, an esoteric "alternative," promoted by critics of mainstream education and not taken seriously as a potential competitor for standard American forms of testing, may soon become a dominant feature of the American educational landscape. Not surprisingly, as states and school districts begin to consider performance assessments as potential official measures of achievement, questions are being raised about the extent to which the new technology of assessment will really be able to deliver reliable, valid, and fair measures of student achievement.

In most discussions of performance assessment, it is tacitly assumed that the new forms of performance assessment are intended to function just as traditional tests do and so can be judged without complication against traditional psychometric criteria. In one sense, this assumption is correct. Assessments used

[1] Preparation of this paper was supported by grants from the U.S. Department of Education for research on the New Standards Project. I want to thank Robert Mislevy for his helpful comments on an earlier draft of this paper. Errors of fact or interpretation, however, are my own.

in officially evaluating students or schools need to deliver information that educators and the public can trust. We must, therefore, have reasonable confidence that the scores offered do not depend unduly on personal biases of judges, special features of the performance tasks that are set for students, or accidents of the conditions under which the assessments are administered. In another sense, however, performance assessment represents such a significant departure from traditional American testing practice that we may never come to grips with either its possibilities or genuine problems unless we address it on its own terms.

There are at least two important ways in which performance assessment as it is developing today differs in fundamental aims and assumptions from our current standardized tests. The first difference, which has been widely acknowledged in general educational discussions but only infrequently discussed when matters of technical adequacy are on the table, is that the new performance assessments are intended to function as integral elements within the education system, rather than as external monitors of the system. The new performance assessments are meant to set standards to which students and teachers can direct their efforts. They must maintain their validity even when they are "taught to." And they must be capable of exemplifying standards, setting clear targets for instruction and learning efforts.

By contrast, our traditional tests are designed mostly to monitor the system. They are not expected to reflect directly curriculum or instructional content. They work as indirect measures—dipsticks or thermometers—of how a student or an institution is doing, and their validity lies substantially in their ability to predict future performances or correlate with more directly observed capabilities. They work best when no one is "teaching the test," for they tend to lose predictive validity when students are drilled on items closely matched to test forms.

The second difference between performance assessments and traditional testing is that they emerge from rather different assumptions about the nature of human knowledge and competence. Traditional testing is rooted in assumptions of associationism, expressed perhaps most elegantly in the psychological writings of one of the founders of American testing, Edward L. Thorndike. Performance assessment, by contrast, is more consonant with the epistemological assumptions of pragmatism, as expressed by John Dewey, George Herbert Mead, and, more recently, by theorists of situated cognition.

Associationist epistemology assumes that knowledge and skill can be fully characterized in terms of collections of separate bits of mental associations or stimulus-response pairs. It further supposes—along with information processing and structuralist theories of human knowing—that competence is fundamentally a function of internally represented knowledge. Associationist theories of testing seek to identify "traits" or abilities that are, in their essence, unrelated to particular contexts of performance. Traditional testing treats context effects as "noise" or "error." This is true even when the language of traits and general cognitive abilities is avoided in favor of defining domains of competence in academic terms (for example, in the ACT tests or in most standardized achievement tests).

Pragmatic epistemology, by contrast, assumes that competence is being able to perform well *in particular environments*. The tools, people, and institutional demands of a situation interact with an individual's state of preparedness to produce a particular performance. Consonant with pragmatic epistemology and theories of situated cognition, performance assessment is focused more on *certifying accomplishments* than on identifying enduring traits of individuals.

Too much can be made of the differences between performance assessment and traditional testing. The two must serve some of the same functions in society: assessing how well students are doing in meeting the educational goals set for them; evaluating how well educational institutions are doing at helping students meet these goals; estimating how well a person is likely to perform in a new environment of study or work. We must develop a robust technology of performance assessment that will allow it to serve these functions. That job is now barely begun. We cannot expect to succeed at this task, however, if we apply unreflectively the technical tools of a measurement technology different in intent and epistemological underpinnings. Performance assessment will require its own tools and technical standards. I hope to begin here a conversation about those tools and standards that I anticipate will need to continue for many years.

Portfolio and Performance: A Social Design Problem

To get started, it will help to build up our image of performance assessment in its own terms, not as a substitute or alternative to current forms of testing, but as a set of procedures designed to serve certain social and institutional functions. To do this, let us imagine that no formal tests or examinations of any kind exist and that schooling as we know it today has not yet been invented. Imagine instead an apprenticeship system in which young people learn their specialties by working in the production shops of local craftsmen and then go into a broader world seeking employment or commissions or further study and training opportunities. These young people, of course, would need to carry with them some evidence of their capabilities, a *bona fides*, as it were, of the likelihood of their producing good work in the future or of learning well and becoming a credit to the institution of advanced study that they joined.

To give our problem some personal reality, imagine that the person under consideration is a young woman who has apprenticed as a weaver in her home village and has heard of new and advanced techniques to be learned in the workshop of a famous weaver in the central town of her region. She goes to that town hoping to gain a place as a senior apprentice in that workshop. What questions might the master weaver in the new, town workshop ask?

First, we can imagine, would come the question, "Do you know how to do the kind of work required here?" To this, our young woman might say, "Let me show you some examples of work I have done." She would open her satchel and arrange the cloth she has brought in a display designed to show it to best advantage. Examining the weavings, the master weaver would decide whether the quality and kind of work displayed was up to the standards she hoped for and contained the range and variety of styles expected of successful entrants to the workshop. This is *portfolio assessment* in its simplest, purest form!

Is that all that is needed? Well, not quite. For how is the master weaver, faced with an unknown young aspirant, to be sure that the work displayed in the young woman's portfolio is really her own? It is not hard to imagine a system of assurances, of *certification,* arising in the region. The young woman's portfolio would, then, include a letter, written in the hand of the craftsman in whose shop she did her initial apprentice work and stamped with the establishment's known seal, certifying that the pieces of work in her portfolio (which might be named and catalogued) were her own. The village craftsman might even add a few words about the reliability and willingness to work—in short, the *character*—of the aspirant. We now have a portfolio that carries a more trustworthy record than just the work itself. We have, in effect, a *certification of accomplishments*.

Of course, this system depends on personal connections or at least personal reputation. Suppose, however, that the master weaver had become so famous that applicants for apprentice places in her workshop came not just from a local region but from distant places. Then she might never have heard of the local weaver in whose shop the applicant's portfolio was presumably assembled. The master weaver might be satisfied with another layer of certification, perhaps from the Regional Association of Weavers from which the applicant came, attesting to the honesty and reliability of the craftsman certifying the young person's work. But if the applicant came from very far away, and if there were applicants from many regions, even that certification might not seem trustworthy. The master weaver might then check applicants' ability to produce work of the kind seen in their portfolios by watching them produce a similar piece of work. Then there could be no question of authenticity of the work. If the quality matched that of the portfolio work, the applicant could be accepted with confidence.

We have arrived at the idea of an *on-demand performance assessment*. Presumably the on-demand performances would be as much like the portfolio items as possible, but there would be some compromises necessary: The on-demand performances would need to be shorter and manageable under controlled conditions. As a result, they would be less likely to indicate originality and flexibility than the full portfolio items. And there might be some forms of weaving whose complexity could not be assessed under the constraints of the on-demand performance. Nevertheless, on-demand performances are a welcome, perhaps even necessary, complement to the portfolio, for besides being practical as a check on portfolio work, the on-demand performances would offer another advantage: a reliable basis for comparing a number of applicants on a common set of tasks.

No doubt the master weaver would not want to base admissions solely on these prescribed performances. They would not, after all, be able to indicate originality, flexibility, or design skill as well as the full portfolio items. But as an "anchor" for interpreting the very varied work likely to show up in a set of individual portfolios, on-demand performances would be very useful. Additionally, the director might decide to give applicants an on-demand *learning* test, perhaps presenting them with a new style of craftsmanship not contained in their portfolios and observing how they did at mastering this new mode of work. Or she might pose an *invention* task, asking the applicants to design and create a weaving that was deliberately different from any included in their portfolios.

Over time, we can imagine, the master weaver would probably become good at judging invented work, as well as the portfolio items and performances on the other on-demand tasks. But if the work load became too great, she might assemble a *jury* to examine portfolios and evaluate performances. Disagreements among members of the jury would be resolved by discussion among the members and, when necessary, intervention by the master.

We have now designed an assessment system that seems to contain all the elements necessary for a reliable and valid, geographically portable credential. With just a bit more elaboration, we can imagine a system that also functioned to increase weaving ability throughout the land. It is likely that in local workshops everywhere, trainers of young weavers would try to communicate to their apprentices the criteria for the kind of work they should put in their portfolios when applying to work with the master weaver. They probably would also give their apprentices some practice in the kinds of on-demand learning and invention tasks they knew the master weaver would ask of her applicants. The portfolio criteria and the known kinds of on-demand tasks would, in effect, establish *standards* for weaving education, clear goals toward which aspiring weavers could work.

Notice what is *not* in this system. There are no tests of "general weaving ability" or of component abilities in the craft. Apprentices might spend some time learning how to do the component skills of weaving—selecting fibers, dyeing them, and the like. But there is no need for any kind of separate demonstration of competence on these components in the portfolio, because it is understood that weavings in a portfolio were done "from scratch." This understanding is carried in the certification and authentication of portfolios of work. Similarly, there may be some "exercise" or "practice" weavings in the course of weaving education; but these needn't enter the portfolio, because they are not considered major accomplishments.

What are the limitations in the picture I have just sketched? "Well," you might say, "your story is nice enough for the very special case you have sketched: a small and elite training workshop, a field of work in which the products can be carried in a small satchel. What about a world in which hundreds of thousands of young people need to be educated in a wide variety of fields and where a master in a field cannot count on personally knowing the certifiers or jurors of another region? Where reliability and fairness in judgment are paramount concerns? Where, unlike crafts or the visual and performing arts, we are not used to defining competence in terms of visible products or performances? And where there is no catalog of established *genres* of work that can be used as a guideline in setting criteria for portfolio entries?"

These are all legitimate questions, and each deserves a thoughtful response. The responses, which are developed in the remainder of this article, will engage us in a consideration of many of the classical issues of measurement technology: reliability of scoring, generalizability of observed performance, and content and construct validity.

Customized Education on a Large Scale

Let me begin with the issue of a small and elite system versus a mass system. In my story, the young woman had the advantages of small institutions and personalized education both in her initial, local apprentice preparation and in the process of applying for a place in the master weaver's workshop. Her apprenticeship teacher worked with her, coached her, evaluated her work, and helped her choose the weaving products she selected for her portfolio. The master weaver's workshop was also small and personalized enough that the Master herself, or a small committee, could evaluate the portfolio of work and judge the necessary on-demand performances.

In American education today, very nearly the opposite situation holds. The schools are serving huge numbers of young people, and they are often large, impersonal places. No one coaches students through the process of preparing a portfolio of accomplishments. Furthermore, the universities and companies students want to join after finishing school often process thousands of applications every year. No single jury could possibly study every portfolio or judge every on-demand performance. But if multiple juries were used, there would be no chance to compare judgments and talk through disagreements. The juries might each go off in a different direction, and there would be no way to ensure common standards of judgment. On the face of it, it does not look as though we could adapt the apprenticeship and portfolio model to our mass education system.

Yet we cannot afford *not* to do so. Everything points to the fact that the mass education system as we know it has to be reworked rather than accepted as a fact of life. Just about every current program of education reform calls for personalized education and small, face-to-face education communities. Everywhere, educators aiming for superior performance are trying to figure out how to break large institutions into smaller units of personal relationship and human accountability. Small schools, schools within schools, vertical teaching teams, and the like are often our beacon lights of reform.

These beacons of hope have, for the most part, been created by working *against* the existing education system. One of the most important challenges in education reform is to create a system that supports, instead of suppresses, personalized and customized education. One feature of such a changed system will have to be a very different method of assessment than the one we use now. The portfolio-cum-on-demand-checkups approach in my story is a good starting place. Can we make it work on a large—but not a *mass*—scale? Can we make it work for many people, without losing its essential personal elements?

Assuring Reliability and Fairness of Judgment

The first problem to be solved is the need for hundreds, probably thousands, of juries. How can common standards of judgment be established and monitored? Suppose a "master jury" is established: a set of people whom everyone respects and who are practiced at reaching agreement with each other. We know a good deal about attaining agreement among judges who are in continuous interaction with one another. But we should anticipate difficulties in getting adequate agreements between *groups* of judges who are not able to engage in face-to-face communication.

The New Standards Project[2] has addressed this problem through a system of *benchmark tasks* that are used to train scorers and to check their reliability from time to time while scoring is underway. Benchmark tasks are pieces of student

[2] The New Standards Project is a partnership of 19 states and 6 urban school districts that are developing shared standards and a system of performance and portfolio assessments to instantiate those standards. New Standards partners either will use the New Standards products directly as part of their state or district assessment programs or will participate in a process of linking their own assessments to the standards established by the partnership.

work selected by a group of lead teachers as exemplifying a certain score level.[3] The benchmark tasks are presented to candidate scorers along with extended commentary explaining why each warrants a particular score. In training, candidate scorers discuss benchmark papers extensively, and candidates remain "in training" until they meet a criterion of assigning scores identical to the benchmark score to 16 of 20 successive papers that they have not seen before. By training scorers to match their judgments to the benchmark papers rather than to one other, it is possible to calibrate different groups of scorers to the same standard. New Standards has found that, using this form of training-to-benchmark procedure, it is possible to maintain scoring reliability even when scoring is done at multiple, dispersed sites. In principle, there does not seem to be any reason this procedure cannot be spread to indefinitely many separate scoring sites. This would make it possible to handle scoring of a very large number of performance assessments by simply expanding the number of scoring juries.

The benchmark papers approach will require some adaptation for portfolios. The important difference is that portfolios will vary from student to student. All students will not have responded to the same question, so there will be no simple way to select a paper that exemplifies a certain level of response. The scorer's job will be not only to judge the quality of a particular piece of work but also to decide whether a collection of portfolio entries, considered as a whole, displays all of the capabilities that are valued. This is a much more sophisticated and demanding judgment task than scoring a single piece of work. Furthermore, for both educational (helping teachers to internalize the standards and providing early feedback to students on how they are doing) and economic reasons, we will probably want a system in which the faculty of a school is charged with the first round of portfolio scoring for its students.

To make this work, we will need a sophisticated set of guidelines and benchmark examples for scoring portfolios. I discuss the nature of these later when considering the question of content and construct validity of assessments. But even with these guidelines and benchmarks available, assuring objective and fair judgments from teachers working in many dispersed sites will require a system in which different groups of judges check one another's scores and are in sufficient communication with one another that, via conversation, challenge, and

[3] New Standards grades student work at five score levels: 1, 2, 3, 4, and 4+ (candidate for honors). A 4 score is described as "meeting the standard." A paper marked 4+ will receive special honors jurying.

argument, they develop and maintain common standards of judgment. The system I have in mind is not very different from the one developed over decades of practice in Great Britain and other countries with decentralized education traditions that use traditional essay examination systems. In the British *moderation* system, each stage of the examining process—establishing the course syllabus, setting the questions, describing criteria for different grades, grading sample (equivalent to our benchmark) papers, the overall distribution of grades—is cross-checked by an individual or a group from a sister institution. In the least formal form of moderation, moderators look over the grades given by assigned graders and confirm their reasonableness. Knowing what we do about "confirmation bias," we will prefer independent rescoring as a more stringent form of moderation.

We have yet to establish an American version of moderation, suitable to the vast size of this country and responsive, too, to our particular political organization in which states retain constitutional authority for education and in which, in some states, all authority over curriculum—and, hence, the content of any form of assessment that is "taught to"—is further delegated to individual school districts. New Standards envisages a multilayered "auditing" system: Initially, scores will be assigned to individual student portfolios by a school faculty—that is, the faculty acting corporately, not as individual teachers. The corporate grading, especially when guided by a set of criteria for portfolios-as-whole and for individual pieces, would act to stabilize scoring and to protect students from the arbitrary judgments of individual teachers.

Next, a selection of every school's portfolios would be sent elsewhere (perhaps to another school, perhaps to a central quality control board) for rescoring. If the rescoring team agreed with the original scores to a sufficient degree, all of the school's scores would be certified, and the faculty grades for all student portfolios would stand. If there were insufficient agreement, a full rescoring might be called for or perhaps only a rescoring of those portfolios on the borderline between "meeting the standard" and not quite meeting it. In New Standards, that would mean rescoring all 3 and 4 portfolios: the former to ensure that students have not been unfairly denied credit for meeting the standard; the latter to control against schools' setting standards that are too lenient.

At the next "layer" of auditing, a state quality board would need to receive reports of the cross-grading and to certify that the school-level auditing is proceeding appropriately, perhaps making some site visits or adding another layer of regrading to do this with confidence. Finally, the New Standards partners want assurance that grading standards are the same from state to state. They are looking to the New Standards Project to add yet another layer of auditing, one that ensures that each partner's auditing system is based on equivalent criteria and that it is operating efficiently and equitably.

Many technical and social issues remain to be resolved in this plan, among them the questions of what proportion of portfolios needs to be regraded and what constitutes a sufficient degree of agreement between original scores and audited scores. Answers to these questions will depend, in part, on what is to be done with the scores. If "high stakes" for individual students are attached—for example, if the scores are to play a role in college admissions or gaining employment—tighter levels of agreement are needed than if no major decisions depend on them. Alternatively or in combination, appropriate systems of appeal allowing students who feel that their work has been unfairly judged to call for a rescoring may remove some of the pressure for near-perfect agreement in original scoring and simultaneously create a sense of visible, public fairness in the system.

As this last comment suggests, the appropriate combination of rescoring criteria and appeal processes is not just a technical matter to be resolved by statisticians and decision theorists. It is at least as much a question of social design: finding a system that people—students, teachers, parents, colleges, and employers—are able to believe in and willing to trust.

Beyond the Satchel: The Problem of "Representative Work"

There is more to obtaining a fair judgment of a student than just assuring that scorers agree with one another. There is also the problem of how to get a fair picture of a student's competence from a few pieces of work. Performance tasks, whether in portfolios or in on-demand assessments, stand for more than just themselves. They are meant to be "representative work" capable of yielding information about the student's competence in a *field* of accomplishment.

Common sense tells us that a single example of a person's work is less good as an indicator of general competence in a domain than a collection of his or her work. But how many exemplars are needed? And how should they relate to one another? These are the classic questions of *generalizability* in measurement.

It is an established fact of mental measurements that any two test items are likely to be only weakly correlated. That means it is not safe to generalize from performance on one item to performance on any other. The problem of generalizability is solved in traditional testing by using many short test items. A score is then created by totaling performance on these items. The score based on 30 or 50 items can, if the items have been well selected, generalize well to a different set of items that have been selected according to a similar set of principles. No single test item carries much weight, but the collection as a whole has some generalizability.

The solution of using many different items will not work for performance assessment for the simple, practical reason that performance assessment tasks take a long time to do, and so it is not possible to administer many of them to any single student. Recent work on generalizability in performance assessment is providing estimates of the number of performance tasks needed to stabilize scores. These studies suggest that 10 or 15 performance tasks in a given domain—not 30 to 50, as previously thought—are needed. But if each task requires about an hour to complete, taking even 10 or 15 is more than we would reasonably ask of an individual student.

Two solutions to this dilemma are typically proposed. The first, available when the score of interest is for a school, a district, or a state but not an individual student, is to use some form of *matrix* or light sampling. Each student takes only a few of the performance tasks, and the results from many students are pooled to yield a score for the group based on many tasks. Because the group score is based on many tasks, it can be adequately generalizable without undue testing time for individual students. How to choose the tasks for a matrix performance assessment and how to distribute them among students are new problems for assessment theory and practice. Research is required using multiple patterns of task administration to yield data on how performances relate to one another.

A special problem for some forms of performance assessment arises from the fact that the tasks may require some whole-class activities. This is so for many of the New Standards mathematics and English language arts tasks. The whole-class activities are sometimes designed to “level the playing field” for the assessment by providing some common experiences for all students before they take the test. Or, these activities may be an integral part of what is to be measured, as would occur when class discussion or other “teamwork” skills were to be assessed. In either case, when whole-class activities are used, it is not possible to give different tasks to different students in the same class; so we cannot derive a classwide group score by matrixing within the classroom.

This in turn may limit the kinds of on-demand testing we can use to derive schoolwide scores. Giving different classes different tasks and then summing across several classes to yield a school score would work only in very large schools. And, even in those schools, there might be interactions between tasks and classes (for example, because teachers emphasized different aspects of the curriculum, or because of different student ability levels in different classrooms) that would make the matrixing procedure invalid. If research shows that these difficulties do in fact hold, a schoolwide score will be possible only if shorter tasks (so more can be administered) or tasks that do not require full-classroom activity (so different students can do different tasks) are used. These patterns and possibilities, too, will have to be worked out over the next several years as performance assessment comes into wider use.

The most promising long-term solution to the generalizability problem, however, probably lies not in constraining on-demand performances to the requirements of generalizability but in breaking down the distinction between learning events and measurement events, so that most measurement information comes as a natural by-product of worthwhile learning activities in which students engage throughout the school year. That is just what portfolio assessment does. This seems to be a happy case in which technical measurement demands coincide with desirable pedagogical practice.

Defining the Domain: Questions of Validity in Performance Assessment

Empirical patterns of association alone cannot provide robust solutions to the generalizability problem. We will also need better ways than we now have of describing the domains over which generalization is expected. We must, in other words, develop principled ways to answer the question, *generalization to what?* A random collection of tasks, no matter how large the number or how elegant the matrix design, cannot represent an individual's or a group's competence in a field of knowledge or skill. The tasks must be systematically related to a careful definition of the field. This requirement takes us into questions of *content* and *construct validity*.

Validity is where performance assessment has its strongest potential. Indeed, the movement toward performance assessment has arisen largely in response to a widespread belief that American standardized tests do a poor job of representing the kind of knowledge and skill that we value. The decomposition of important knowledge and skill into disconnected bits and the decontextualization from meaningful situations of use that standardized tests impose virtually ensure their inability to validly assess complex capabilities in which knowledge and skill are combined to produce meaningful intellectual, artistic, or design products. By contrast, the performance assessments and portfolio projects now coming into use have won accolades in many quarters for their capacity to represent the kinds of knowledge and skills most educators hope will become the dominant focus of teaching and learning in the future.

Until now, however, these accolades have been based mainly on inspection of individual performance tasks and portfolio entries. Many of these are elegant and appear to do a fair job of representing the new forms of academic content that educators in mathematics, science, English, and other disciplines value. But how do they represent, as a collection, the range of knowledge and skill we expect of competent students? To ask this question is to inquire about the *construct validity* of an assessment.

There is no way to establish the construct validity of a collection of tasks in the absence of an agreed-upon framework that describes the knowledge and skill that students are expected to learn and that should be sampled by the assessment. In the U.S., we actively avoided developing such frameworks until just a few years ago, for reasons linked to our historical commitment to local control of education. The absence of such frameworks has made it essentially impossible to deal sensibly with the problem of content or construct validity in assessment. There has been no way to establish what the content of assessment tasks *should* be or how to interpret the collection of tasks as representative of a domain.

A notable exception with respect to content validity is the College Board's Advanced Placement (AP) program, which is a *syllabus-based* assessment system. Schools that want to prepare their students for the AP exam in any particular year receive a syllabus telling them what the exam will cover. The syllabus guides text selection, teaching, and in-class paper writing and testing. Assuming that students have been in a course that largely follows the syllabus, the exam they take at the end of the year is, by definition, content valid. However, the AP program does not explicitly take on questions of construct validity. A student is assigned a score by summing across the different sections of content, but there is no specification of how to make inferences about what the student knows *about the domain as a whole.*

The new movement to develop consensus *content standards* in the major school subject matters represents an important, indeed crucial, step forward in defining content-valid testing. The movement is furthest advanced in the field of mathematics, where the National Council of Teachers of Mathematics (NCTM) *Standards for School Mathematics* has led the way. In attempting to build specifications for the New Standards assessment program in mathematics, we have found that the NCTM *Standards* are helpful in providing criteria for judging whether individual tasks are content valid: that is, whether they reflect knowledge and skills defined in the *Standards*. The *Standards* thus provide a grounding for judgments of content validity, although they are not nearly as precise as the Advanced Placement syllabi.

It is not an accident that the NCTM *Standards*, as such, do not provide a principled basis for making judgments of *construct* validity. They are not sufficiently constraining, and they do not specify how the various elements of content are related to one another. By intent, the *Standards* lay out a very broad field of aspiration for mathematics education and do not specify exactly what any school should teach. As a result, schools attempting to use the *Standards* to guide their curriculum redesign have found that they have to make many difficult choices about what to emphasize and what to exclude or to make optional with guidance from the *Standards*. We have had the identical problem in trying to use the *Standards* directly in designing New Standards assessments.

The solution that appears to be workable is to develop what we have come to call a *Framework for Balance,* which takes up where the broad national consensus standards leave off. The *Framework* dimensionalizes the content standards and specifies which dimensions *must* and which *may* be included in the assessment. These dimensions include strands of specific knowledge (e.g., probability, fractions and decimals, constructive geometry) as well as skills such as displaying data, problem solving, graphing, and manipulating equations. The *Framework* further specifies the broad *genres* of student work that the assessment program as a whole should sample (e.g., a survey study, a physical experiment). Tasks or extended projects falling within the same genre can be sensibly scored using the same rubric; different genres require different scoring rubrics. Thus, the criteria for excellent work are genre-specific.

The *Framework for Balance* takes an important step toward the specifications that will be needed for establishing the construct validity of assessments. However, the *Framework for Balance* does not—by design—specify the precise performance tasks or portfolio projects that must be included in the assessment. Instead, New Standards is developing a process by which a *collection* of student work (an individual's work in a portfolio; a group's work in the collection of matrixed performance tasks) can be *mapped* to the dimensions of the framework to show how the work taken as a whole demonstrates competence in all of the dimensions specified. This should allow us to identify many different specific assessment packages—and thus many different specific curricula—that conform to the framework and allow principled judgments about students' competence in a subject matter.

In Conclusion

I opened this essay with the claim that performance assessment is designed for a different set of social functions than traditional American testing and that it is grounded in a different set of epistemological assumptions. I think it is fair to say that the social design requirements for an assessment system that can set targets for educational effort are today the driving force in assessment research. Because we have decided we need new forms of assessment, many groups are at work developing them. Typically, the time and ingenuity needed to solve the practical problems—of scoring, of educating teachers in the new methods, of generating new assessment tasks, of managing large-scale operations while retaining personalization—absorb most of the resources of assessment development groups. There is not much energy left for reflection on the theoretical aspects of what they are doing.

Yet this practical work is not without theoretical significance. In the spirit of pragmatic epistemology, the efforts to create a new technology of assessment are beginning to point the way toward a new theory. Efforts to define performance standards are producing candidate definitions of fields of accomplishment. The challenge of developing techniques for mapping tasks and performances to frameworks is likely to refine these definitions quickly. And the idea of *genres* as defining classes of performance situations may be a first step toward a cognitive theory of situations.

One thing that is now clear is that performance assessment cannot develop on solid ground without much more explicit theories of situated cognition than are now available. We need ways of defining situations in terms of their cognitive demands and opportunities so that we can begin to develop a cognitive theory of accomplishment. A cognitive theory of accomplishment would explain how situation and person interact to produce a competent performance, rather than looking for traits that are stable across contexts, or, alternatively, contexts that override personal characteristics. Pragmatic philosophy called for this kind of interactionist theory of cognition. Now so too does the practical demand for new forms of assessment.