DOCUMENT RESUME

ED 390 904 TM 024 217

AUTHOR	Bunderson, C. Victor
TITLE	Measurement Science and Training.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-88-63
PUB DATE	88
NOTE	84p.
PUB TYPE	Viewpoints (Opinion/Position Papers, Essays, etc.) (120)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Artificial Intelligence; *Cognitive Psychology; Economic Factors; *Educational Assessment; *Evaluation Methods; Interaction; *Measurement Techniques; *Organizational Change; *Performance Based Assessment; *Training
IDENTIFIERS	Mastery Evaluation

ABSTRACT

The need for training and retraining is a central element in current discussions about the economy of the United States. This paper is designed to introduce training practitioners to some new concepts about how measurement science can provide a new framework for assessing progress and can add new discipline to the development, implementation, and conduct of training. The paper is intended to be a discussion-focusing chapter in a forthcoming book sponsored by the American Society of Training and Development in which other chapters will be written by training practitioners. The paper demonstrates that measurement science, revitalized by interactive technologies and the disciplines of cognitive science, instructional science, applied artificial intelligence, and studies of organizational change, can be used to address the economic challenge our nation currently faces. Basic concepts of measurement science as applied to training and performance are described, and newer concepts, such as mastery assessment systems and the four generations of computerized measurement, are explored to demonstrate their applicability to training and performance measurement. A hypothetical future training application, using the advantages offered by advances in measurement science, is described. (Contains 1 table and 23 references.) (Author)

RR-88-63

RESEARCH REPORT

MEASUREMENT SCIENCE AND TRAINING

C. Victor Bunderson

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
November 1988

MEASUREMENT SCIENCE AND TRAINING

by

C. Victor Bunderson[1]

May 1988

[1] This paper will appear as a chapter in Improving Human Resource Development through Measurement, a forthcoming book from the American Society of Training and Development.

Abstract

The need for training and retraining is a central element in current discussions about the economy of the United States. This paper was designed to introduce training practitioners to some new concepts about how measurement science can provide a new framework for assessing progress and can add new discip ine to the development, implementation, and conduct of training. The paper is a discussion-focusing chapter in a forthcoming ASTD-sponsored book in which other chapters are written by training practitioners. The paper demonstrates that measurement science, revitalized by interactive technologies and the disciplines of cognitive science, instructional science, applied artificial intelligence, and studies of organizational change, can be used to address the economic challenge our nation currently faces. Basic concepts of measurement science as applied to training and performance are described, and newer concepts, such as mastery assessment systems and the four generations of computerized measurement, are explored to demonstrate their applicability to training and performance measurement. A hypothetical future training application, making use of the advantages offered by advances in measurement science, is described.

Ca ı a revitalized measurement science help solve the nation's training problems? The need for training and retraining is currently a central element in discussions about the economy of the United States. Our country's ability to compete in world markets is being challenged severely by the willingness of workers in other countries to work for lower wages, and by their increasing ability to mass produce quality goods. As a result, our economy's ability to maintain a high standard of living for all of our people is increasingly doubtful. Many companies with jobs requiring minimal education and little ability to respond flexibly have automated or sought offshore labor. The jobs that remain require higher levels of technical knowledge and skills, yet our pool of educated and skilled people is growing smaller. A large, poorly educated underclass, which includes many minorities, is increasing. Not only is our standard of living undermined through decreased economic capacity, but the democratic foundations of our society are threatened because democracy requires an educated citizenry.

We must educate the young and prepare them for increasingly complex and technical roles in a world economy. We must retrain adults, on the average, eight times during their careers. High levels of competence are needed. As pointed out in the recent report by the Carnegie Forum on Education and the Economy (1986), the alternative to the old model of using semiskilled labor is to "revise our view of the role of the worker in the economy. In the future, high-wage level societies will be those whose economies are based on the use on a wide scale of very highly skilled workers, backed up by the most advanced technologies available" (p. 13).

Alan Greenspan (1988), Chairman of the Federal Reserve Board of Governors, spoke this year about America's ability to function in a global economy. He stressed the intellectual component as key to the rise of the GNP since the turn of the century and certainly the key to any future rise: "In addition to modernizing the physical capital stock, we must also concentrate on broadening and deepening our 'human capital'" (p. 8).

He also stated,

> I am referring not only to those [skills] that are specific to particular jobs and t us can quickly become obsolete. Rather, I want to stress the need to acquire broad analytical and problem-solving capabilities that will facilitate the processing of information and enhance one's ability to adapt to the demands of a complex, dynamic economy (p. 9).

The issue addressed in this paper is whether a revitalized measurement science can be of significant help in achieving new levels of effectiveness in the kinds of training we will need to be competitive in the world. The science of educational measurement deals with that elusive, valuable, and invisible quality--human competence. Measurement scientists have struggled with the problem of measuring higher levels of competence, including analytical and problem-solving skills. One product of this struggle has been tests given under standardized conditions with the purpose of making visible that hard-to-observe but highly valuable resource, high-level competence. However, measurement science as it has functioned in the past is not likely to make a significant impact on the problem of achieving high levels of technical competence through retraining the American work force. One reason is that it has not dealt with measuring learning and improvement but rather with predicting future events from current scores. Another is that the conventional models of expertise have been too simplistic. Measurement science itself is now in the throes of change in response to new interactive technologies and the impact of new scientific disciplines. These emerging changes have the potential to revitalize measurement science and make it more relevant to training practice.

The new technological alternatives that affect the development of human competence include interactive training delivery systems. These delivery systems employ computers interfaced to mass memories containing large files of video images, audio messages, computer-generated graphics, and computer-generated sound and text. The systems provide not only a wide variety of formats for information display, but also new response modes as alternatives to the standard multiple-choice format. These include keyed responses employing natural language and mathematical symbols, touch screens, and the use of cursor keys or a mouse to point to screen locations and "move" objects from one place to another on the screen. Interactive training delivery systems, however, have not reached their full potential. The reasons include cost, lack of wide distribution and organizational inertia, and, most fundamentally, an inadequate understanding of learning processes and teaching processes.

Better scientific foundations can help reduce the problems that have slowed the use of interactive training systems. Increased scientific understanding is currently emerging from new disciplines such as the cognitive sciences, instructional science, applied artificial intelligence, and studies of organizational change. Measurement science is currently undergoing a revitalization through these disciplines and the use of new delivery systems. At Educational Testing Service, we are eagerly engaged in this revitalization process. The Research Division at ETS has scientists involved in all of the five areas mentioned above and in the use of new interactive delivery systems. This paper will emphasize the interplay between training and a measurement science revitalized by new scientific methods and technologies. In addition to discussing how measurement science can improve training, it will also examine how the problems and perspectives provided by interactive training systems are stretching measurement science in new ways and producing a cross-fertilization which should greatly expand both disciplines. The following discussions embrace (1) measurement science concepts and definitions, and (2) a training scenario for the future in which these concepts are applied.

Concepts of Measurement Science

Two measurement science definitions useful for training professionals are educational measurement and performance measurement. Educational measurement is the process of specifying the positions, for educational purposes, of persons, situations, or events on educationally relevant scales under stipulated conditions. A related concept is performance measurement, which is a similar process of specifying the positions of persons, situations, or events on scales relevant to valuable (generally economically valuable) accomplishments under stipulated conditions.

Solutions to training problems will have to be delivered by an applied science of training. To what extent can this applied science be improved by better measurement? Measurement scientists must rise to a new set of challenges in integrating measurement into training and improved performance in a way that has not been accomplished to date.

Common Applications Of Measurement In Training And Performance

Admissions, Selection, Certification and Licensing

To date, the more sophisticated contributions of measurement science have not been an intimate part of training or education. Instead, the most common applications of measurement science include the use of carefully crafted admissions tests for selection of applicants into colleges or work assignments, and the use of testing for certification or licensing in an occupation. Combinations of placement and diagnostic tests have been given before studies begin, and proficiency tests have been administered midway and at the end of a course. However, measurement science has not yet made a powerful approach to measuring the growth of human competence as it develops over time as a result of learning.

Counseling and Guidance

Tests of ability, interest, personality, and other attributes are currently a substantial part of the process of counseling and guidance, whether it be for career planning, college admissions, employment counseling, or therapeutic purposes. These tests will continue to serve important needs, but they do not directly impact training.

Performance Engineering

The "educational measurement community" consists of professionals in the traditions of Thorndike, Thurstone, and others who have focused on measuring aptitudes and knowledge, and on the psychometric and statistical issues of scaling, equating of scales, validity, reliability, and so forth. They constitute a group that is very different from the professionals who have focused on performance engineering. This group developed out of the fields of human factors (earlier, "human engineering") and job training. While the initial focus of human engineering was to design equipment to be compatible with the capabilities and limits of people, the field has broadened to include designing jobs and specifying the needed human skills. Performance engineering now impacts both selection and training as well as the design of hardware and software systems.

Measurement is an integral part of performance engineering. The 656-page textbook commissioned by Bell Laboratories in 1978, Human Performance Engineering: A Guide for System Designers (Bailey, 1982), is replete with examples of the use of measurement in assessing human performance globally, as it relates to specific jobs and tasks and, specifically, as it relates to human limits and differences in sensing, perceiving, responding, processing, problem solving, decision making, and so forth. Gilbert (1978), in his book Human Competence, uses performance measurement as a fundamental instrument for producing high-level competence. His methods can be used for developing models of performance with economic consequences, for evaluating the payoff of competence, and for determining the cost of lack of competence. The applied fields of performance engineering have not as yet found the advanced statistical and psychometric models of educational measurement necessary for their work, nor have they given much emphasis to the validity of inferences from scores. By seeking the common goal of achieving higher levels of competence, the sister fields of educational measurement and performance engineering may each benefit greatly from the findings and methods of the other.

Formative Evaluation

Applied measurement has been used in formative evaluation for a variety of training-related purposes. In formative evaluation, data are used to improve a system as it is developed or implemented. Surveys are often used to obtain information during the development of training programs and survey data can help to refine job analysis and to determine the most critical tasks to be addressed in a training program. Item analysis can be used to improve evolving training programs through formative testing of exercise units. Clearly, measurement methods could make a much larger contribution to the formative evaluation of training in the future.

Summative Evaluation

Summative evaluation uses data to make an overall judgment about the value of a program. Policy makers frequently seek summative evaluations of training programs either as a way of proving that the programs are achieving good results, or as a way of attacking current training policies. When policy makers have positions at stake, the validity of inferences made from summative measurements is frequently strained. It behooves measurement science to offer scales and guidelines that can be used to support valid inferences and interpretations. Measurement experts also need to be persistently persuasive in encouraging users to meet the conditions for the proper use and interpretation of measures.

In summary, educational measurement has made its contributions at the beginning and end of learning activities, but not toward improving the ongoing process of learning. Applied fields like performance engineering and interactive training are making important contributions to the processes of human learning, but they should be challenged to consider measurement issues more seriously.

Validity Is The Foundation Concept Of Measurement Science

When measurement scientists ask whether what is supposed to measured is in fact being measured, they are addressing the issue of validity. The validity of inferences from measures of competence is a key issue in training. This is particularly true of inferences about the status of trainee growth and what can be done to smooth the paths toward their higher levels of competence. Interactive technologies (such as computer- controlled videodiscs) offer promise for providing job-like simulations that can be used to measure performance during training. Unfortunately, most uses of these technologies have depended primarily on face validity--the inference that the resemblance of a simulation to an actual job means that performance in the simulation is equivalent to performance on the job. Technology-based performance measures will be strengthened by the evidence of criterion-related validity, whereby the performance levels demonstrated in the simulation are validated against actual on-the-job criteria. For example, criterion-related validity is present in flight simulators because simulator training is followed by actual flying hours. Unfortunately, many training simulations are not validated against real job-performance criteria.

As important as criterion-related validity is, construct validation may be the greatest contribution that measurement science can offer to training and performance. The term construct validity refers to the accuracy of inferences that are based on an understanding of the fundamental constructs (ideas derived from a model or theory) of the domain or task. Scales generated from test items (through factor analysis, for example) can be useful in understanding the dimensionality of the domain. If four scales are needed to deal with the tasks in the domain, the constructs that explain what each dimension represents may lead to inferences about how to train more effectively. The constructs that explain growing competence, as an individual progresses from novice to expert, are potentially the most powerful contributions of measurement science. While these constructs can be developed from cognitive or instructional science, they must be validated by measurement science.

The road to true mastery is seldom smooth or continuous. There are discontinuous changes in the cognitive processes and the mental models used by novices, advanced beginners, adepts, and experts. If measurement can highlight and reveal the nature of these mental models, this information can then be used to refine the specific presentations and feedback in a program of instruction and can potentially speed learning.

Scales, Construct Dimensions, And Anchors: An Example

A major contribution of measurement science has been the development of psychometric and statistical models. These, in turn, have led to scales that span wide ranges of human proficiency. One example is the adult literacy scales of the National Assessment of Educational Progress (NAEP) (Kirsch and Jungeblut, 1986). In this study, the domain of adult literacy was sampled by using simulation tasks that resembled what adults do in the real world. These adult-literacy items were administered one-on-one by trained observers to a representative sample of 3600 young adults, 17 to 24 years of age. The subsequent analysis produced three unidimensional scales (Kirsch, 1987):

Prose Literacy--knowledge and skills needed to understand and use information from texts that include editorials, news stories, poems, and the like;

Document Literacy--the knowledge and skills needed to locate and use information contained in job applications, payroll forms, bus schedules, maps, tables, indexes, and so forth;

Quantitative Literacy--the knowledge and skills needed to apply, either sequentially or alone, arithmetic operations embedded in printed materials, such as balancing a checkbook, figuring out a tip, completing an order form, or determining the amount of interest from a loan advertisement.

The discussion which follows was adapted from Kirsch. Tasks that are representative of these three types of literacy were scaled using item response theory (IRT). IRT is a mathematical model for estimating the probability that a person will respond correctly to a particular task. To determine this probability, analyses within a given scale were carried out in two steps. First, the parameters of the tasks were estimated. For the NAEP, these parameters include item discrimination, item difficulty, and, where appropriate, guessing. Second, levels of proficiency were estimated for individuals and groups. The tasks, which anchor points along the scales, provide a criterion-referenced interpretation of the corresponding points. The group proficiency estimates provide data for norm-referenced interpretations (Kirsch, 1987).

To aid in interpreting performance at various levels along each scale, supplemental information was developed by selecting benchmark tasks (also called anchors) and identifying underlying characteristics that contribute to task difficulty. For example, tasks on the document scale seemed to vary as a function of three characteristics: (a) the number of features or categories of information that the reader had to locate or use, (b) the relationship of the wording in the question to the wording in the document, and (c) the number of distractors or potentially correct answers in the material.

These analyses suggest that literacy tasks often require the reader to apply complex information-processing skills and strategies to a variety of linguistic structures. Awareness of these complexities deepens our understanding of the nature of literacy. Furthermore, it replaces overly simplistic labels like "illiterate" or "functionally illiterate" with task-referenced levels of performance and associated information-processing skills and strategies.

Moreover, gaining a better understanding of the underlying skills and strategies associated with performance at various levels on each of the scales has important implications, not only for building more reliable and valid literacy measures, but also for designing instructional programs. Adult literacy programs are too often based on models suitable for elementary schools, where reading is, for the most part, restricted to the use of narrative texts. Results from the NAEP study suggest that emphasis on a single aspect of literacy may not lead to the acquisition of the skills and strategies needed to fully participate in our society.

Quantitative models of measurement science, in the form of item response theory and its variations, made the NAEP analysis possible. The scientific fruits of this large-scale measurement activity are continuing, as Educational Testing Service, which administers and analyzes the national assessment, has made a substantial investment toward finding ways to use the NAEP scales and constructs to improve adult literacy. The goal is to develop systems of integrated assessment and instruction embodying Continuous Measurement, a new generation of computerized measurement explained below.

Concepts Of Measurement Science Expanded By New Technologies

Recent developments in computerized measurement can be summarized by placing them in a four-generation framework (Bunderson, Inouye, & Olsen, in press). Each generation represents an application of measurement science through technologies of increasing sophistication and power. The generations are:

Generation 1, Computerized Testing (CT): Administering conventional tests by computer.

Generation 2, Computerized Adaptive Testing (CAT): Adapting the test to the individual test-taker by selecting each succeeding task on the basis of the test-taker's performance on previous tasks.

Generation 3, Continuous Measurement (CM): Using calibrated measures embedded in a curriculum to continuously and unobtrusively estimate changes in each learner's proficiency.

Generation 4, Intelligent Measurement (IM): Introducing knowledge-based (artificially intelligent) computing to the decision making processes of computerized measurement.

The First Generation, Computerized Tests, CT

The CT generation is typified by translations of existing t3sts to computerized formats, or by new, non-adaptive tests which are similar to manually administered tests but utilize computer capabilities for all or most test administration processes.

The first generation offers many advantages to the processes of measurement. Not the least of these are immediate scoring, immediate feedback, and testing on demand. CT generally proceeds faster than pencil and paper test formats. New response formats are made possible through computerized testing, including typed responses in the forms of numbers, equations, words, or phrases; pointing as a response through the use of a touch screen, cursor keys, or mouse; and auditory responses. Joy sticks, track balls, steering wheels, and so on offer other possible modes of response. Many new displays are also available, including motion pictures, animation, randomly accessed audio, computer-generated overlay on video images, etc. The computer makes it easy to measure the latencies of individual responses and elapsed times for complex sequences of actions. It is also possible to vary the length of time that different displays are presented. Data collected from computerized tests can be transmitted electronically to a central site for statistical analysis, calibration, scale equating, and various analyses aimed at test improvement.

CT enables first-generation tests to be more job-like and, hence, less decontextualized. Such tests may be less formidable to the test taker, not only because they can be constructed to resemble familiar things in the work situation but also because computerized test items appear one at a time, not as a formidable booklet filled with page after page of test questions.

The Second Generation, Computerized Adaptive Testing, CAT

The CAT generation of computerized educational measurement is typified by computer-administered tests in which the presentation of a next task, or the decision to stop presenting new tasks, is adaptive to the individual test taker. The presentation of each new task is determined by calculations based on the test-taker's performance on previous tasks. A task may be an item or a complex standardized situation involving one or more responses.

CAT tests can adapt on the basis of several types of information. The most common adaptation is based on an estimated scale value. That is, on the basis of previous responses, an estimation of the position of the learner is made on the underlying latent trait scale. (A latent trait is both the statistically defined scale and the inferred construct explaining score differences.) This estimate guides the selection of the next item, which may be easier or more difficult. CAT tests generally use item response theory, which permits both test items and test takers to be identified by their positions on the same underlying latent trait scale. As the estimate of the test taker's position becomes more certain, items closer to the estimated scale value can be presented to the learner, thus making the estimate of the learner's position on the scale increasingly precise. An accurate estimate of the test taker's position on the scale may be obtained with from 30 to 60 percent fewer items than are required with a paper and pencil test, leading to further reductions in testing time over CT.

Computer adaptive tests are excellent for identifying the exact scale position of a learner on an underlying latent-trait scale. They are not ideal for determining whether a learner is above or below the cut point used for determining certification, licensing, or admission into a program, because these testing applications do not require the exact scale position of each test taker. Also, the content specifications of the test may be violated if only a small subset of items are administered. Another form of adaptive test, called the Computerized Mastery Test, has been developed by Educational Testing Service, to optimize cut-score decisions adaptively. This test adaptively selects well-specified item clusters to sample the content specification adsquately and to decide quickly whether the test taker's score is above or below the cut score.

Computerized tests may adapt with regard to response time by varying the speed of presentation or by requiring faster and faster responses. Such tests may converge on the individual test taker's characteristic speed of perception or speed of response under stipulated conditions.

Test content and the sequence of displays may also be adapted, as in a simulation. Flight simulations, for example, are adaptive in that the parameters of the simulation change in response to the trainee's engagement of the simulation controls.

Construct-valid scoring schemes will be developed as scientists define the constructs that operate in given simulations and relate them to the underlying processes. These scoring schemes will produce multiple scores from the same simulation to inform learners and instructors about the learner's performance and how it may differ from the performance of more expert simulation operators.

The Third Generation, Continuous Measurement, CM

The continuous measurement generation (CM) uses calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in each learner's proficiency. The tasks used as the units of measurement may be items, item clusters, exercises, unit tests, or independent work assignments. Changes may be observed in the amount learned, in proficiency on different tasks, in the trajectory through the domain, and in the student's profile as a learner.

CM is the first generation of educational measurement that holds the potential to be fully responsive to the needs of learning and instruction. CM is integrated into an instructional system to provide continuing feedback for use by both learners and instructors. It is designed to guide and facilitate growth on several underlying mastery scales. These mastery scales have the properties described earlier for the adult literacy scales. They always span all, or a coherent part, of a domain of expertise. For example, adult literacy was spanned by the domain of simulation items selected for the NAEP Young Adult Literacy Assessment. Tl·· domain of interest may be collapsed into a few key constructs, as with the prose literacy, document literacy, and quantitative literacy constructs described earlier. These constructs are anchored by demonstrable performance on reference tasks along the way to mastery, so that context and specificity can be given to positions on the scale, rather than simply confronting the learner with a numerical score. Indeed, numbers do not have to be communicated to the learners and instructors--just the scales, the different types of tasks on each scale, and the complexities associated with mastering the constructs that lead to success on the tasks.

Continuous measurement is unobtrusive. It is not segregated from learning, as are mid-terms, finals, or periodic tests in training courses. Instead of imposing obtrusive and formidable forms of testing, continuous measurement resembles that friendly form of testing we have all experienced when our teachers and fellow students discuss correct and incorrect answers on a very recent quiz so that all can learn from the experience. It is diagnostic and associated with advice to help each learner achieve a desired goal.

Continuous measurement is a key component of an integrated system of instruction and assessment. Mastery Assessment System is one name given to the assessment component of an integrated instructional system that allows us to estimate the learners' positions, track their trajectories on the underlying scales, and make inferences about the deeper constructs fundamental to those scales. It is discussed in detail later.

The Fourth Generation, Intelligent Measurement, IM

Intelligent Measurement (IM) is defined as the application of knowledge-based computing to any of the subprocesses of educational or performance measurement.

Intelligent measurement introduces knowledge based (artificially intelligent) computing to the decision making processes of computerized measurement. The potential applications of this capability are virtually unlimited. Three applications of immediate interest are (1) the intelligent scoring of constructed responses, (2) the intelligent interpretations of score profiles, and (3) generating intelligent prescriptions for appropriate instruction in the continuous measurement environment.

Computerized scoring of constructed responses would be a major boon to measurement. Currently, labor-intensive methods of hulistic scoring are employed for grading writing exercises, pieces of art work, and architectural drawings. Human judges grade the productions after training in a set of standards illustrated and anchored by ample productions representing various points on the rating scale. For example, six essays might be provided, anchoring the attributes of ratings from one to six. Researchers have examined the possibility of computerized grading of computer programs written in the Pascal language (Bennett et al., 1988). The Advanced Placement program for computer science requires the ability to write such programs, and many are submitted on floppy disks. Bennett and his colleagues are working with Elliot Soloway of Yale University to examine the applicability of his artificially intelligent program called PROUST (Johnson & Soloway, 1987) to the scoring of these programs. Although there is promise in these methods, early applications are not imminent.

IM can also be used to obtain intelligent interpretation of score profiles. In this application, the expertise of human counselors, who advise individuals based on their score profiles, is captured by knowledge engineers and placed in a knowledge base so that the computer may go through the same reasoning as an experienced counselor and generate similar interpretations. Such expertise in intelligently interpreting the profiles can then be replicated at many locations.

In the most complex application of intelligent measurement, the computer would generate prescriptions, hints, or prompts to advise learners who are moving through a system employing continuous measurement. Scores from multiple tasks along the way could be obtained to provide a profile of characteristics of the individual learner as well as the learner's trajectory on the underlying scales. Human experience with such technologically advanced systems must be gained before this expertise can be captured and used in intelligent continuous measurement.

Concepts Of Mastery Assessment Systems[1]

The concept of mastery assessment systems, a CM generation concept, was developed by ETS researchers during 1986 and 1987. Two multi-year research projects have been initiated to develop mastery assessment systems, one in adult literacy and the other in middle-school science (Lipson, 1988).

Two features of mastery assessment systems may be noted at the outset. First, a mastery system is a part of a larger learning/ performance information system for a specified curriculum, but it is not a curriculum in and of itself. Second, the term "mastery" does not refer to minimum competence alone.

A Mastery System is Part of a Learning/Performance Information System.

Learning/performance information system is a term used to describe the management system which coordinates curriculum materials, instructional resources and schedules, and mastery assessment components into a single, integrated system. The mastery system is the measurement component of the learning/performance information system. It provides assessment information about how learners or trainees progress on their paths from novice to expert and how their performances compare with the exemplary performance we would attribute to "masters" of the subject assessed. The total volume of instructional materials, management data, and applications software which occupy the learning/performance information system data banks may be much larger than the mastery assessment system materials. The learning/performance information system itself is housed in a networked set of work stations which communicate with a central file server for data management. This system may be operated primarily for training purposes in a laboratory or learning assistance center, or the system may be more work oriented. It may be used by a work group in a local area network to automate activities. In this application, the mastery assessment materials and the curriculum materials comprise a lesser part of the system, which is devoted primarily to supporting the work to be automated by the group.

Two functions of a mastery system within curriculum. While a mastery assessment system is not itself a curriculum, it provides two important functions when integrated with a curriculum. The word "curriculum" is ill-defined. It may be used at different times to refer to the historical evolution of instructional content and courses, to all of the courses in a catalogue, or to the instruction materials used in a single course of study. The Latin root of the word "curriculum" means to run, as to run a course. A training course is designed to run the learners along an instructor-selected path through some domain of knowledge and expertise. The domain itself is larger than any course, which is limited to a finite period of time; thus, multiple courses, at different levels of expertise, are common. In addition, there may not be adequate coverage of a domain by any particular course, nor practice which leads to real mastery of the tasks performed by domain experts. This possible disconnect between the course and the domain represents one need a mastery system might fill. Another is to mark progress by providing milestones with performance standards so that the learners' progress may be demonstrated and appropriate actions taken along the way.

When embedded within a curriculum, or more properly, within a specific course, a Mastery Assessment System provides two important functions in relation to the curriculum:

1) Through a "mastery map," it represents to trainees and supervisors a view of the domain that may be broader than the particular course. In this way, it helps to establish the representativeness and appropriateness to the domain (as known by experts) of the elements selected to be mastered in a particular course.

2) It enables instructors and learners to determine the levels and standards of achievement along many possible paths to mastery. These are communicated not by numerical abstractions, but by anchors and reference tasks that refer to different levels of growth and to achievements in the actual world of work.

Development of a mastery assessment system involves an interdisciplinary mix of measurement, cognitive, and instructional sciences. Since the domain represented may be broader than any particular course, the mastery map must be customized, just as an instructor selects chapters from textbooks.

A Mastery Assessment System Focuses on Mastery Beyond Minimum Competence.

When a measurement organization obtains group consensus on learning milestones in a particular subject area, that consensus usually converges on what might be called a minimum competence standard. The mastery that is assessed in a mastery system, on the other hand, looks forward to a time when, after long commitment and effort, the learner has obtained a life-long capability. It is important, therefore, in developing mastery assessment systems to identify human beings who are truly masters or exemplars of the performance to be achieved by the end of the training.

Mastery signifies achievement of a high order of performance goals. Mastery is personal and unique and is achieved after long periods of persistence and commitment. Performance engineering can identify the potentials for improvement in well defined tasks and teach these performances to others. However, the nation needs high levels of flexible performance capability in our workers. The term mastery reflects this need by going beyond exemplary performance in a single, well prescribed job. Training programs must aim toward development of mastery because jobs change rapidly, and workers must be prepared to adapt to changing demands. The assessment of higher levels of mastery often involves examining unique productions (e.g., complex problem solving, oral presentations, written analyses, portfolios, etc.). Some of the precursors of mastery can be assessed at earlier levels of growth by encouraging the learner to practice some element of mastery appropriate to the growth stage. At intermediate stages of learning, the trainee can thus experience the rewards of persistance and the ability to expand upon what has been learned until able to do something extremely well.

Assessment includes the use of standardized measures of competence and guidance for judging the precursors of mastery at various levels. This guidance may include disciplined subjective scoring or, in the future, intelligent computerized scoring. Instructionally sensitive assessment will have new properties and paradigms that have not yet been fully developed by a measurement science built to support certification, selection, and classification.

Components of a Mastery System

A mastery assessment system will require a learning information system as described above. The earlier discussion of computerized measurement assumes that the most effective and complete examples of mastery assessment systems will be implemented in heavily computerized learning assistance centers or in work areas with local area networks. Simpler forms of mastery assessment systems that could be implemented partly on paper could also be developed, as well as transition technologies that begin with paper delivery and progress to the heavily computerized model presented in this paper. For the simplest system, at least one computer should be available. It will provide scoring and record keeping, and serve as a primitive learning information system.

Major non-hardware components of a mastery system include the following:

1. A mastery map usable by learners and instructors to envision and communicate learning goals.
2. Reference tasks.
3. Calibration of items and reference tasks.
4. An instruction-oriented scoring system for each reference task.
5. A professional development program to help instructors/supervisors learn to use the system effectively.

These components are intended to serve instruction and to be linked with instructional components. Instructional components will include repeated practice in reference tasks, subscoring to guide the instructor in coaching, and report-generating systems for both learners and instructors. Coaching, in this context, is analogous to the instructional process used by an excellent athletic coach. The process may include modelling the desired performance, observing practice trials, prompting, encouraging, and fading the prompts as the performance becomes adequate.

The Mastery Map. Each training enterprise requires that a domain be mapped and defined in terms of knowledge elements, job elements and the kinds of tasks that novices, adepts, and masters perform. The mastery map is a way of making a domain visible. It gives the trainees and their supervisors or instructors an overview of the journey as learning begins. The mastery map also permits communication about initial placement and the next steps for each individual trainee. The mastery map could be visualized on a large wall display for all, but individual maps with status information would be made available graphically on computer screens, hierarchically organized and shown a screen at a time.

A key feature of the mastery map is information about the current status of the learner or trainee. A colorful way to make progress is to turn sections of the map red, yellow, or green, depending on whether a person has passed through a challenge successfully (green), needs further coaching (yellow), or has encountered such difficulty as to indicate a lack of readiness for that challenge (red).

The term "mastery learning" is sometimes used to describe instruction that demands rigid adherence to a linear sequence of tests of minimum competence with no choice but to pass each test before going on. However, the mastery system concept presented here promotes a more expansive (and more historically accurate) view of mastery. It describes a human master as one who has attained exceptional levels of competence, not the narrow definition implied by "mastering a test." In fact, as different paths are appropriate for different learners, the system does not demand that challenges be passed in a prescribed order, or that every task be passed by every learner.

The mastery assessment system also offers different approaches to delivery and implementation.

Reference tasks. A reference task is generally more complex than a single item. It may be a testlet (a related group of items) (Wainer & Kiely, 1987), a curriculum-embedded exercise requiring multiple responses, or a simulation exercise. A reference task is contextualized. It refers to some real-world work that communicates the relevance of what is practiced to intelligent lay people. The scoring of a reference task may also refer to component process constructs important to exemplary performance on the task and useful in coaching. A record of an individual's progress on reference tasks can serve to build self-confidence. In this way, it may serve as a reference example--a benchmark to look back to when attempting new challenges. Table 1 contrasts test items and reference tasks.

Insert Table 1 about here

Calibration. Reference tasks can be placed on scales to show the degree of growth they represent. Test items, perhaps grouped into clusters or testlets, can also be placed on such scales. For example, the following tasks were used in the NAEP study of the literacy of young adults (Kirsch & Jungeblut, 1986). They assess literacy skills used in interpreting documents (forms, maps, charts, etc.). The scale values are statistically determined measures of difficulty based on item response theory. In the literacy study, they are also described and explained in terms of the task features that account for variations in difficulty.

Sign your name on the line that reads "signature."

Scale value: 110

Put an x on the map where two particular streets intersect.

Scale value: 249

Fill in a check to pay a particular credit card bill.

Scale value: 259

Use a bus schedule to answer: On Saturday morning, what time does the second bus arrive at the downtown terminal?

Scale value: 334

Use a bus schedule to answer: On Saturday afternoon, if you miss the 2:35 bus leaving Hancock and Buena Ventura going to Flintridge and Academy, how long will you have to wait for the next bus?

Scale value: 365

These examples illustrate the calibration of reference tasks in a mastery system. Initially, calibrated values could reflect both the experts' judgment and early empirical results. Later calibrations with large numbers of learners would be based on a new generation of appropriate statistical models and would become more accurate and stable. Once the reference tasks in a set are calibrated, the scale values are given meaning by showing the constructs of knowledge and skill required to succeed at tasks with a given range of values. Each mastery system has its own calibration. The scale and scale interpretation are developed and validated for a particular content, level, and purpose.

Instruction-related scoring of reference tasks. Test items are normally scored dichotomously--correct and incorrect. It is possible to score reference tasks more finely to connect performance with instructional strategy. Instruction-related scoring allows us to identify common or especially troublesome misconceptions and erroneous procedures and to provide appropriate feedback.

Reference tasks may be very close to the criterion of real-world performance. These reference tasks offer a means to measure critical dimensions of performance, such as speed, accuracy, and strategy. In this way, the reference task becomes more like the actual criterion than a predictor of later performance. Sometimes computer simulations are "better than the real thing" because we cannot instrument and measure the real thing in situations that are, for example, extremely hazardous or costly, or that pose other formidable difficulties.

A professional development program for supervisors and instructors. A mastery system is always accompanied by a professional development program for supervisors, who often will be the instructors. Professional development includes training in the appropriate use and interpretation of measures and in methods to build and maintain appropriate climates for teaching and learning. The learning information system provides a tracking system so that instructors can make professional decisions about how to manage the paths of a group and its individual members. More advanced systems will provide information to guide the coaching of individuals and groups. Instructors will have to learn new practices in relation to these technological tools so that they can be successful at using the mastery system for placement, tracking, training management, and ultimately for individual and small group diagnosis and coaching.

A mastery system is designed to function within a community of learners or workers, and users must learn to build and sustain such environments. A mastery system provides the opportunity to build and support a cooperative community whose goals are to help and encourage one another, to teach one another, and to facilitate the maximum amount of learning for all.

How Can a Technology-Extended Measurement Science Benefit Training and Performance?

Answers to this question can be explored within the framework of Instructional Systems Development (ISD). The steps in ISD are frequently divided into five categories (see Branson & Grow, 1987):

1. Front-end analysis.
2. Design activities.
3. Development activities.
4. Implementation.
5. Continuing evaluation.

In this part of the paper, the section headings correspond to all but stage two of these five stages of ISD: front-end analysis, instructional development (featuring prototypes that use data for iterative improvement), implementation, and continuing evaluation (emphasizing ongoing data collection for continually controlling and improving the interactive training system). Except for the impact of construct valid scales to represent the domain and construct-related scoring within reference tasks, there will be no further discussion of possible impacts of measurement science on design activities. Before going into ISD stages, we will describe a scenario for a future training application to illustrate how training can be enriched by technology and measurement science.

A Description Of A Future Training Application

This scenario describes installation of a new delivery system involving interactive technology with integrated training and measurement. The system is installed in a company to solve a large-scale training problem and the scenario includes a transition plan for moving from the current training environment to a new technology- and measurement-intensive environment.

The Application

Assume that a large company operating at numerous locations has decided to install computer work stations, networked into clusters for functional work groups of 10 or more workers. The software to be implemented includes some fairly complicated mail-order data processing software which is being introduced slowly to replace manual procedures. This new software is integrated with word processing, spreadsheets, and other productivity software packages so that a clear career path is possible for clerical workers. The goal is for many trainees to start at the bottom, learning basic and repetitive aspects of the mail-order system, and for some of them to move up to more advanced secretarial and administrative positions. The company has experienced substantial turnover at the secretarial and clerical levels.

The Talent Pool

Because turnover is high and the available labor pool is not well prepared, the company finds that it must dip into the less well educated levels of the labor pool where there are literacy problems. The available candidates are not illiterate, but their literacy skills are frequently insufficient for keeping up with the training classes the company has instituted. In addition, the company's proposed software for automating the manual mail-order processes has proven to be extremely training-critical. That is, substantial increases in productivity can be obtained by exemplary performers, but the use of automated systems by those with lower literacy skills or without proper training has led to numerous and costly mistakes. The company is not anxious to implement automation widely or rapidly until the implementation can progress without these mistakes. In this scenario, the company has installed the clusters of microcomputers in a small number of its most exemplary work groups and has made some initial attempts at training old and new workers.

The Current Training Activity

The personnel department has established a series of classes and a "lab room" in which machines are installed for practice on the software packages. It also has an agreement with local community colleges to provide literacy classes where needed. The process of training in the lab room is low-key in that there are not enough trainers to expand it to the work force at large, nor is the laboratory adequate to support many trainees. Management would like to decentralize the training activity to the locations where the machines will be installed. Because of the current training constraints, management has slowed the rate of machine installation.

Management's Plan for Introducing a Distributed Measurement-Intensive Training Environment

The plan has several steps:

1. Conducting an intensive front-end analysis using new concepts for defining and placing a value on expertise.
2. Developing a mastery map, reference tasks, and scoring procedures based on front-end analysis.
3. Instrumenting the existing classes for continuing measurement.
4. Selecting and adapting existing interactive training materials and developing new materials.
5. Iterative testing and revision in a small number of test sites.
6. Implementation plans and tactics for replication to multiple

sites.

Conducting an intensive front-end analysis. The company managers first want to make clearly visible the level of expertise they hope to achieve. They want both themselves and the project personnel, including the trainees, to see a visual representation of the mastery to be obtained and to be able to mark individual and group progress visibly, through status information on a mastery map. In order to mark trainee progress, they want reference tasks to be identified and measures developed to assess performance on all reference tasks during the course toward mastery. Management commissions a front-end analysis that will yield these products, as well as the other products of a more conventional front-end analysis. As in the past, they expect the analysis to yield job tasks and objective performance levels, but they want more than verbal descriptions. They want samples of the products, and data on the speed and accuracy with which work products can and should be produced (work standards). They want diagrams of the procedures followed by more expert workers. They want a list of the names of the most expert workers who have been located, within or outside the company, so that these people can both serve as models and provide data on exemplary performance.

Management also wants an economic analysis of the payoff from a target level of expertise. They wish to locate or develop a quantitative model demonstrating the benefits of expertise and the costs of errors, mistakes, multiple revisions and slowness due to lack of expertise. They know that as one alternative they can commission a performance audit, following the methods outlined by Thomas Gilbert (1978). A performance audit yields the ratio of the performance of exemplary performers on tasks and subtasks to the performance of typical, merely adept, performers. These ratios are the Potentials for Improving Performance (PIPS). The performance audit identifies which PIPs have the greatest potential and how these are related to costs and economic benefits. A team is commissioned to seek this type of economic information from vendors or, if the information is not available, to recommend an action plan and a budget by which the company may pursue these questions to the depth desired.

The front-end analysis will also determine the beginning performance levels of individuals from the labor pool, including the ranges of literacy available and the minimum keyboarding, productivity, and computer use skills. A "job literacy" analysis is essential to determine the levels required for success in the job training and on the job. Consistent with the needs and with the capabilities available in the pool, analysts seek to determine the literacy standards needed to reach exemplary job performance levels. Management may then commission appropriate literacy instruction so that more applicants from the available pool can benefit from job training.

Developing the mastery map and the reference tasks. After the front-end analysis is available, the project team develops the initial version of the mastery map and a set (initially sparse) of reference tasks. Fortunately, a commercial package is found which provides speed and error scores with norms for the word processing and related keyboarding skills. The company producing this package is commissioned to produce assessments based on reference tasks to define increasing levels of expertise in the use of the mail-order processing software. This company has authoring software appropriate to keyboard oriented jobs and is willing to produce rapidly the new assessment tasks for the order-processing software.

Instrumenting the existing classes for continuing measurement. The performance measurement systems are installed in the work station laboratory so that at certain points during practice and learning, as well as at the end point, the proficiency of the trainees can be measured. The wall chart versions of the mastery maps are installed in both classroom and laboratory and the initial computerized mastery maps are installed on the computers.

Paper and pencil tests and other group-administered measurement activities are also introduced into the current classes to help get a better assessment of how the trainees are progressing from day to day. It is planned that this instrumentation in the laboratory and the classroom will produce data that can be used to incrementally improve both of these training resources and to guide the implementation of the distributed training environment.

Selecting and adapting training materials. The developers look for available interactive training material so that the training can be decentralized from the training room to the work units. They find some commercial packages for the office productivity software but require development of reference tasks identified on the mastery map but not covered by the commercial lessons. The developers use "lean development methods" by producing things rapidly for presentation by the instructors in the training room and putting on the computer only those materials that have been tried out with the trainees and have shown their ability to communicate clearly. The new reference tasks, used initially only for measurement, provide files of response data from which lists of common errors are obtained. Feedback messages are prepared and reference tasks are slowly improved by providing better and more thorough feedback. The training will be designed primarily for interactive administration once the system is more fully developed. As an interactive, on-demand system, it will continue to be used after implementation, under the direction of the supervisors, as turnover brings in new workers.

Iterative testing and revision. While the instrumentation for iterative testing is being installed, a schedule is worked out for the implementation of the initial pilot test work stations. The pilot test work groups are identified and equipment is ordered for them. A training schedule is established to combine training room activities with practice at individual work stations using real work assignments. The general strategy is to use the networked computer work stations in addition to the practice equipment in the training room. Thus the second stage of the implementation plan is developed. It builds on what is being learned in the practice lab and transfers it to a plan for the work units.

Time for extensive revision is scheduled between each of the installations of the software and training materials into a pilot-test work group. In this manner it is hoped that the interactive practice materials, the on-line measurements, and the implementation plan can all be improved.

During the pilot testing, great care is taken to develop and refine an implementation plan. As the testing indicates that supervisors are the key to successful implementation, they are brought to one of the training classes and taught how to tutor individual workers in the practice laboratory. It is also found that using the identified expert workers is an extremely effective tactic in the implementation process. Their speed and error statistics are made available, and they come to the pilot-test sites to talk to the workers, answer their questions, and provide encouragement. As new workers achieve high levels of proficiency and become "masters," they also begin to provide coaching and encouragement for others, both in their own work groups and beyond.

Implementation throughout the company. The implementation strategy allowed for three iterations in pilot-test work sites before full scale implementation was made throughout the company. Not all reference tasks were available until the third pilot test, so that each iteration involved a more complete system. Not only was there a desire and effort to ensure that the training and interactive measurement materials were soundly conceived, there was also concern that all of the human management elements involved in the implementation of this technology be explored prior to full-scale implementation.

A Description of the Mature System

Company management found that it took 16 months for development and pilot testing and another 6 months for full-scale implementation. During this process, they received continual feedback on the levels of mastery achieved by trainees as a result of the instrumentation installed in each work group. The reports were used by line supervisors and were well understood by the top management group. The economic analyses for the program convinced management of the value of paying full attention to the development of human expertise. Models of the cost of not having the needed worker expertise thoroughly justified this investment in human resource development. The models were refined continually during the implementation process, and the performance measures were revised twice a year to make sure that they were current and had not become familiar to trainees.

Management was very pleased with the successful implementation of this project, but they knew that the job was not finished. Not only had they not reached the top levels of exemplary expertise defined before the project began, but technology had moved ahead during the 22 months required for development and implementation. It was possible to identify individuals who had broken through the ceilings defined by mastery levels established in the front-end analysis. These "star performers" had exceeded the previous standards for exemplary performance through a combination of improved software productivity tools and higher levels of human expertise.

Management wanted to continue the growth which had occurred by setting clear standards for the growth of mastery and having a measurement system to monitor it. But they did not wish to make the tactical error of measuring individual performances in a hidden or manipulative manner. Therefore, they requested only aggregated data from each work group and established quality circles in which workers could study the data themselves and try to raise their own performance levels. In so doing, each work group (including the supervisor) was rewarded both for increases in productivity and for providing suggestions for improvements in the total human development system.

After implementing a continuing evaluation and improvement cycle, delegating it to the workers involved, and providing a reward system so that workers could share in the benefits of the productivity gains, management observed a continuing upward trend in productivity and expertise statistics. When work groups made requests for additional hardware and software, management had a framework for evaluating the economic tradeoffs of installing it and providing the on-the-job training needed.

Summary

This scenario allows us to single out several features of the proposed measurement- and technology-intensive approach.

1) Management used a front-end analysis that clearly identified high levels of human expertise with real people, not just standardized descriptions of job elements and tasks.

2) Economic costs and benefits were a part of the front-end analysis: there was a serious attempt to relate human expertise to its economic payoff and lack of expertise to its consequential costs.

3) Management did not develop a one-shot plan a priori, announce it with fanfare, and then distract the whole company with a crash program to automate. They moved slowly, reducing their costs and risks at each step. They first introduced change to

a few pilot-test sites and identified problems through the use of test data. They depended on good training developers to solve the problems identified by testing while the problems were small and contained. This unusually enlightened (and admittedly fictitious) management group did not expect perfection the first time but planned, instead, for problems in implementation. They provided the time and means to identify and correct the problems.

4) The management group took direct control of the human development process and used its training and development resources as a line rather than staff function in implementing productivity improvement through technology and human resource development.

5) The management of implementation was seen to be more important than the development of effective training, which was seen to be more important than the purchase of appropriate hardware.

6) Sound management of human resources through the continuing collection of performance data was not seen as a tactic only for the implementation of an initial change, but was regarded as a tool for continuing improvement.

7) The use of data to improve performance was delegated to the workers. A climate of trust was sought where growth in expertise and productivity was pursued for the good of both the company and individual workers.

Multidisciplinary Approaches To Front-end Analysis

In the scenario, the idealized management group employed a kind of front-end analysis that is an important emerging multidisciplinary technology. This type of front-end analysis cannot be described solely in terms of traditional job and task analyses, although these tested methods (McCormick, 1979; Fine & Wiley, 1971) are important and valuable tools.

Front-end Analysis: A Problem in Representation

We owe to our colleagues in the field of artificial intelligence a better understanding of the power of different representations in shifting information processing from something extremely difficult to something easy; sometimes trivially easy. Selecting the wrong or the right representation for organizing the data and symbolizing the important relationships for a computing problem makes the difference. Bruner (1968), in his essay on instructional theory, argued that the idea of different ways of representing information to human learners can make a very large difference in the difficulty of learning or solving problems. A basic concept in the selection of different representations is that their economy and power is maximized only in relation to a given purpose. Different purposes determine which representations will be awkward and which economical and powerful.

The usual product of front-end analysis--lists of tasks, knowledge areas, and objectives--is used by training developers to produce a syllabus and to guide the design of instructional materials. It has been used by test makers to produce specifications for test items. Its purpose has not been to produce a top level guide to the structure and content of a domain of expertise. Such a guide would maximally communicate about the domain to learners and instructors and provide a way of tracking progress from novice to higher levels. Both mastery maps and reference tasks as defined in Part I can serve as representations suited to these trainee-centered purposes. The revitalized and broadened form of front-end analysis we seek will sensitively select different representational forms for the different purposes of a front-end analysis.

Job tasks and knowledge areas are currently represented to the analyst and questionnaire respondent alike in written words. This form of representation is best for describing the knowledge areas of work. Procedures can be described with words, but they are then described at a knowledge level, which is not the same as the tacit expertise of a true master, who is able to perform procedures effortlessly at the proper time with an expertise level that is "in the fingers and in the gut." These tacit components of job knowledge are often associated with the highest level of expertise, which has become automatized in the human expert. Tacit components are usually missed when the representation is solely verbal. Polanyi (1962) provides a discussion of tacit knowledge.

Therefore, new methods of modelling expertise should be investigated. By "modelling," we mean creating a visible representation of otherwise invisible expertise. Words and numbers are not the only form of representation. Some of these new methods are being pioneered through a branch of applied artificial intelligence called "knowledge acquisition," the most difficult part of knowledge engineering (Hart, 1986; Bunderson, 1987). Knowledge acquisition means acquiring knowledge from one or more human experts and representing it in a publicly visible and usable form, usually symbolic.

A knowledge engineer uses other forms of representation besides words and emphasizes problem solving, not just declarative knowledge (terms, facts, categories). In this way, knowledge engineers penetrate to a higher level of expert knowledge. A knowledge engineer thinks in terms of a knowledge base of formal rules, a data base of facts, and a set of informal heuristics or rules of thumb that experts use to guide the use of rules and facts. (A good heuristic is a guideline that tells the expert where to search for a solution and where not to spend time or effort.) The rules are stated in a formal language such as predicate logic. This representation is more formal than standard written English. Despite this, if left at the level of a representation on paper, such knowledge is still only a symbolic formalization. It is an unusual formalization, however, because it can be tested. The rules in the knowledge base and the facts in the data base may be exercised by programming a prototype expert system and testing it by submitting cases and problems to see if it will solve them as effectively as experts solve them. Producing a complete expert system is obviously too costly for the ordinary front-end analysis, but many problems in knowledge specification can be discovered early through attempts at formalization and through the discipline of testing the formalization (manually, without a computer) using sample cases that the expert is supposed to be able to handle. Consideration of work behaviors to this depth, that is, to the level of the rules and heuristics that experts use in doing their work, is a discipline which has not yet been reduced to practice in job analysis.

The selection of test cases is of utmost importance in testing an expert system. By analogy, the selection of "reference tasks," which refer to the work the experts really have to perform, becomes absolutely critical. These reference tasks become the focus for the development of practice tasks to be used during training and proficiency testing.

Reference Tasks

A good set of reference tasks, like the test cases for assessing the quality of an expert system, will span the range of high-level expertise. These tasks define the upper reaches of the domain. Unlike test cases in an expert system, however, reference tasks must also be developed to span the lower ranges of expertise--novice, advanced beginner, intermediate, adept, etc. These levels of reference tasks are needed in a complete mastery map which tracks the trainee from lower to higher levels of expertise. Selecting a set of reference tasks is an exercise in assessing the representativeness, completeness, and appropriateness of the domain coverage. It is one aspect of the validation process for an assessment and instructional system.

Challenges to Measurement Science

There are many challenges to educational and performance measurement arising from the need and opportunity to improve education and training dramatically. The international economic challenge makes it vital that we succeed in this interdisciplinary endeavor. The challenges to measurement science can be considered as different aspects of the processes of establishing validity.

1) Validating the representations of the domain: the mastery map and the lists of job tasks and knowledge areas. Are they representative and complete? Does the mastery map communicate clearly?

2) Validating the scale constructs derived from the analysis of the tasks in the domain. How many scale dimensions are sufficient to characterize the domain at the top level of the mastery map?

3) Validating the construct-oriented scoring of different reference tasks on different scales; for example, do scores or

states achieved in a simulation task reflect actual expertise in the simulated job?

4) Validating the cognitive and instructional constructs. Do subscores that reflect cognitive constructs, when used in instructional feedback and advice, lead to performance gains?

5) Validating the implementation plan. Does it work? How can it be improved? More generally, what principles of implementation have cross-situational applicability?

The problem of using reference tasks to measure growth presents an immediate challenge to measurement science. It is in this area that measurement science must stretch itself in new directions to create instructionally sensitive assessment. Unfortunately, much of the measurement work developed with test items, which are usually presented at the beginning or end of training, is not applicable to the problems of measuring growth "in process" using reference tasks.

Individual test items are ideal for quickly sampling large domains of knowledge: facts, terms, concepts and simple applications. But short, single items are not well adapted for many kinds of job elements. For example, multiple-choice items must include the right answer, which may be recognized or found by elimination, rather than being recalled or generated. Interactive computer systems can simulate many kinds of job elements not well assessed by short, objective items. They provide a way of administering the reference tasks that introduce much more of the job context into the training and testing situation. Researchers are attempting to define this building block of measurement, the reference task, and to find how it differs from single test items or clusters of test items.

Cognitive science is an important element of this enterprise because cognitive science helps tell us what is important to measure. Instructional science helps us use the scores to provide informative feedback to guide students as they reorganize their knowledge and skills so that they can perform at the next higher level of expertise. As cognitive and instructional science produce a new set of constructs, measurement scientists are then challenged to model these constructs through scores that can be shown to be construct valid. That is, variations in the scores reflect variations in the underlying constructs.

These new constructs must be instructionally sensitive. Older measurement constructs, like aptitudes, abilities, and generalized traits, have not proven useful in guiding instructional activities. Studies of aptitude by treatment interactions have identified individual traits that could lead to prescriptions for alternate instructional treatments. However, the constructs of greatest promise for providing instructional feedback are the cognitive and instructionally oriented constructs that must be derived from deeper analyses of specific tasks. In time, measurement scientists will be challenged to model constructs that cut across a variety of tasks. These are sometimes referred to as "metacognitive skills."

Measurement science is being severely tested by these demands of the real requirements of training. To develop scoring procedures using reference tasks instead of individual test items requires a new look at standard measurement models. To array the reference tasks on a mastery map with a scale of proficiency going from novice to expert requires new foundations for the calibration of tasks. In existing measurement science there are simplifying assumptions that the ability measured is fixed, rather than changing, and that one task has no influence on another. The opposite assumptions are made in training: ability is changing as rapidly as possible as a consequence of practice on the tasks, and one task transfers positively to another. Measurement scientists have generally been only marginally familiar with the concepts and methods of training, instructional systems development, and instructional science. Broadening the contributions of measurement to training requires that measurement professionals learn of the problems and successful principles in the training field. The payoff for this learning may be great, because mastery maps with calibrated reference tasks that can provide continuous measurement both during and after training could potentially reduce costs and improve productivity enough to more than justify the effort.

Tacit knowledge. Knowledge engineering is only one approach to making expertise visible. It still produces written formalizations, even though they are written in a more precise and formal code to be implemented on a computer. True expertise is not limited to knowledge of facts, procedures, rules, and heuristics. It includes tacit components and human motivational components. One example is "role commitment," which refers to the commitment to exert effort toward a new role--to practice over the long periods of time necessary to truly master an area. Role commitment implies, among other things, persistence of effort.

There are other subtle human dimensions in the work environment. It is important to include real people in a front-end analysis and to use them to model and motivate during the implementation and training process. Bandura (1977) has developed a theory of social learning that uses human models as a central element. The human models carry and communicate tacit knowledge and role commitment. As new people enter the elite group of experts, they can be added to the training resource pool and honored as a reward for their effort and as an incentive to others. Also, because much of the expert's performance is never reduced to written or symbolic formalizations, human experts can often notice ways that the trainees can improve and can coach them when given a chance to observe and do so.

Summary: Toward Advances in Front-end Analysis

1) The practice of job and task analysis is being expanded by

multidisciplinary approaches that introduce methods from knowledge engineering, cognitive science, and the social/behavioral sciences.

2) Measurement methods using sampling and data collection can enhance the relevance and completeness of front-end analyses and its sensitivity to subgroup differences.

3) Mastery maps are a suggested new product of front-end analysis. They fill the purpose of modelling the domain, organizing the reference tasks, and tracking movement through the domain for the benefit of trainees and instructors.

4) Analysis methods must be expanded beyond representations that use only written descriptions of work behaviors. Also important are formal representations of critical facts, rules, and heuristics. Tacit human elements are vital, but often can only be conveyed from one person to another through a modelling, coaching process.

5) The selection of appropriate reference tasks at different levels of proficiency is a critical part of front-end analysis. The reference tasks refer to real-world jobs, and to the simpler tasks that novices, intermediates, etc., can perform. They are contextualized, unlike most test items or individual knowledge topics.

6) A revitalized front-end analysis can lead to more informed development of needed training materials. The domain specification leads to a mastery map. The reference tasks

lead to practice exercises and are the building blocks for providing both proficiency testing and individualized feedback.

Using Data To Improve Instructional System Development

Bootstrapping: A Lean Development Approach

The scenario described a "lean development approach" in which the new instructional system was bootstrapped by building on and adding incrementally to an old one. The scenario did not provide all of the details for accomplishing this. Some of the methods used were introduced in the last section. Front-end analysis can be used to produce a mastery map that represents the domain to be mastered and provides a description of the knowledge areas and reference tasks to be dealt with at different stages of development from novice to expert. Imposing this mastery map on the existing training can reveal deficiencies in its coverage. Missing elements may be added either through interactive lessons on the computers themselves, through lessons taught by the instructor in a stand-up mode, or through printed texts.

The mastery map can serve as a management structure for the development and testing of lessons as well as for guiding trainees and instructors. As reference tasks are developed they are checked off. The training situation provides a lean approximation of the integration of tasks and lessons into one course of study. It is "lean" because most of the context is provided by the instructor, who also provides feedback and guidance for slower learners. Interactive lessons have to stand alone, aided only by the supervisor or fellow workers at the work site. They must carry the needed context and provide sufficient feedback for slower learners. In lean development the training center is used and measurements are taken from it as the interactive lessons are developed for distribution to other sites.

During their development, reference tasks may be used to enrich instruction. Later they can be used in the distributed locations. An ideal reference task is developed in such a form that it allows measurement to occur, but provides for repeated practice. In word processing, for example, a reference task may consist of typing a certain kind of document using certain word processing features. It is easy to generate new practice tasks for this kind of exercise. The trick is to provide the scoring so that proficiency can be made visible to the trainees and instructors in terms of some overall scale, so that the scale relates to norms for a larger group, and so that individual feedback messages can be generated.

As computer ir eractive reference tasks with built-in scoring are introduced into training, they enrich ongoing classes and bring closer the day for distribution to remote sites.

Smoothing the Path up the Mastery Mountain using Difficulty Statistics

If a trainee is given a task too easy or too difficult for the trainee's current level of progression, practice is less than optimal and may even be detrimental. It may prove to be either boring or discouraging. Therefore, it is important to know the relative difficulty of reference tasks. Which reference tasks should be presented to novices, to intermediates, and so forth? Measurement can aid in this process through developing models that use data to assess the relative difficulty of each reference task on some underlying proficiency scale or scales.

Very often this empirical process can reveal "cliffs of difficulty" in the sequence of practice tasks. These are the points at which trainees begin to have difficulty, fail to make continuing progress, and encounter frustration. Providing time for iteration, as was suggested in the scenario, allows developers either to present their reference tasks in a different, more accessible manner, or to build a set of scaffolding exercises that can lead the trainee more slowly up the cliff of difficulty. Sometimes tasks are presented in an inefficient and ineffective sequence that creates unnecessary slopes of extraordinary difficulty for many. The use of rough difficulty statistics can reveal this condition and can lead to suggestions for better sequences of reference task exercises, based on a clearer understanding of what is going on at the difficult zone.

Cognitive science may reveal that a cliff of difficulty represents the need for an underlying reorganization of cognitive structures. Expertise is not a smoothly continuous process that grows task by task. Higher levels of expertise are often reached only by restructuring the knowledge, skills, or procedures of a domain in some important ways. Cognitive science provides tools for investigating the constructs involved in the reference tasks found around these cliffs of difficulty. Complexity in underlying structures makes it difficult to scale and calibrate reference tasks at these ranges of discontinuity. Cognitive science uncovers the structure of knowledge and expertise and the restructuring that successful learners generally accomplish. This may make more precise measurement models possible, while rough initial measurement could be used to focus the costly cognitive analysis in the important transition zones. Instructional science provides guidance for developing effective training presentations and informative feedback methods that will be optimally effective in guiding trainees through these difficult paths toward restructured knowledge.

Neither measurement science nor the disciplines that have contributed to training systems development have converged on a standard set of methods for scoring and calibrating reference tasks so that the advantages discussed in thıs section can be achieved. Much cooperation between cognitive scientisťs and measurement experts is necessary before this will be possible. For example, measurement science is still in the process of developing good models for scoring reference tasks based on underlying cognitive structures and processes. Cognitive science lacks good case studies of the cognitive analysis of restructuring that must occur at the points of discontinuity.

Summary: Impact of Measurement Perspectives on Instructional Development

The use of the mastery map as a management structure and the use of iterative testing and revision in a development approach are not especially new ideas. They represent a productive interplay between good measurement and iterative instructional improvement. They are used less frequently than is warranted by their potential payoff. Training development practice can further be enhanced by emerging contributions from measurement science, including determining the relative difficulty of reference tasks and providing the construct-related scoring of such tasks, tied to instructionally effective feedback.

Toward A Design Science Of Implementation

The scenario conveyed a carefully planned process of implementing change slowly and iteratively throughout a whole company. It suggested that the implementation process be directed by line management as a fundamental responsibility for the management of change. The process should involve measurement to test each step of the way and to validate an implementation plan before extending the plan to the whole company.

Anyone familiar with the attempt to improve productivity, whether through training or in job aids, is well aware that the majority of the variance in success or failure is attributable to implementation. The best job-aid technology, the best training plan, can fail if it is implemented improperly. Implementation, and the design of implementation plans, strategies, and tactics for given situations, is worthy of recognition as a new interdiscipline in its own right.

The term <u>design science</u> used in the heading to this section is based on the usage of Herbert Simon (1981), who in his provocative and prophetic book <u>The Sciences of the Artificial</u> defined a science of design as

> a body of intellectually tough, analytic, partly formalizable, partly empirical, teachable doctrine about the design process."
>
> "....Such a science of design not only is possible but is actually emerging at the present time. It has already begun to penetrate the engineering schools, particularly through programs in computer science and "systems engineering," and business schools through management science... We can already see enough of its shape to predict some of the important ways in which engineering schools of tomorrow will differ from departments of physics, and business schools from departments of economics and psychology (p. 132).

The search for sets of principles to guide the process of design is general to all stages of Instructional Systems Development, not just the design stage. Prescriptive principles can guide the process of front-end analysis in seeking and collecting those representations of knowledge that will serve a variety of design and development purposes. Principles will help to establish the content and construct validity of the domain representation and will guide the designing of mastery maps, the selection of reference tasks, and the development of feedback messages.

The implementation process has generally been viewed as a practical concern for managers, rather than a scientific activity. This may, in part, account for the high frequency of failures in implementing new systems and programs. Research-based principles for the design and execution of implementation plans must be validated and shown to produce results. The principles we seek are clearly goal dependent; hence they are better sought within the framework of a design science than a descriptive science, as designs always involve human purposes.

At ETS we have organized a new group of researchers in the Division of Educational Policy Research. We call this group the "Implementation Policy Group." Their job is to investigate implementations of new instructional and assessment systems and to seek general prescriptive principles for introducing technology-intensive applications similar to the one described in the scenario. They seek to discover a set of generalizable principles that can be used to guide the preparation of implementation plans, strategies, and tactics.

A set of principles for guiding the design of an implementation plan will include contributions from a variety of fields. Engineering and measurement are two of the fields that can contribute to the development of automated performance measurement systems and systems for formative purposes. Lean development approaches that provide a bridging strategy from current practice to new practice will also be important. The principles of organizational behavior and other social sciences will be needed to define methods to change the goals, roles, and tools of work groups as technological change is introduced.

In the scenario, the supervisors had to adopt new instructional roles, but new learning and working roles were offered to the workers as well. First. a longer career ladder was introduced, extending from the lower clerical levels to the higher secretarial and administrative assistant levels. Also, workers and supervisors were challenged to improve productivity themselves at the end of the training implementation.

Administering effective implementation policy will require patience. Goals, roles, and traditions do not change readily. Good management, good instruction, good modelling on the part of those who are committed and can demonstrate success--all these will be necessary to change the human parts of a culture as new technology is introduced. John Naisbitt (1984) had a term for it--"high touch." As he pointed out, high tech will be rejected unless there is a large dose of high touch--the human element--to smooth the transition.

Continuing Evaluation

Summative evaluation is the most common concept of evaluation in the minds of policy makers. Moreover, the policy makers in a company are happy to forego any kind of evaluation unless their opponents demand "scientific proof" that (a) a certain program is a success, or (b) a certain program should be dismantled or replaced.

Through the concepts of calibrated scales, with reference tasks as anchors, measurement science can provide evidence to assist policy makers in resolving legitimate concerns about the effectiveness of costly programs. It is possible to link the reference tasks from some well-defined training domain into a set of national norms for the scales that constitute that domain. Subsequent progress of groups of trainees on these scales can then be interpreted in relation to the national norms, as with the three adult literacy scales originally developed and calibrated based on a nationally representative sample of young adults. These scales can be used to show progress in a variety of ways. Statistical methods for equating scales that include common anchor tasks have been developed and refined by large measurement organizations. Calibrated scales of mastery can provide policy makers with progress information and year-to-year trend information.

The NAEP adult literacy scales have the potential to become a part of mastery assessment systems associated with job training systems in a number of areas. By going to the expense of equating the three adult literacy scales to the national scale, policy makers would be able to observe improvements in the general literacy skills of the trainees on a national scale, and improvements in job-specific competence through other scales developed for that purpose.

Care must be taken to support the validity of inferences made from gain scores derived from a proficiency scale. (A gain score is calculated by subtracting the beginning score from the final score.) The use of gain scores often creates serious measurement problems. Mastery scales anchored with reference tasks offer a promising alternative to old-fashioned gain scores. The results of an evaluation need not be reported in terms of numerical values at all. Since the higher-level reference tasks refer to (and it is hoped, predict) actual work competence, evidence that some number of learners are able to perform some difficult reference tasks well, and before instruction could only perform simple tasks, is enough. Experts will already have agreed that such tasks are very job-like and demanding, and such evidence is better than a decontextualized gain score.

Beyond Summative Evaluation

In the scenario, the concept of formative evaluation was demonstrated from the first intervention to the end. Evaluation was implemented as soon as possible into the ongoing classes and laboratory room, by introducing the mastery map and the initial small set of reference tasks. Measurement was used to improve incrementally the different components of the system. A front-end analysis will usually show areas in the training that are not covered by either class or laboratory instruction. New tasks can then be added to the instruction with confidence, to be taught initially by the existing instructors. New practice and feedback elements may be added after each pilot test until the reference tasks and their associated, instructionally-oriented feedbacks approach the goal of standing completely alone. When that goal is reached, the supervisor and fellow workers could guide a new worker through a series of reference tasks and lessons. The reference tasks would simultaneously enable the trainees to demonstrate new skills and progress and provide practice and instructional feedback. The supervisor and workers would still provide much of the context, the motivation, and the instructional management. They would do this as a part of the ongoing work situation. Even after an implementation is complete, a new technology- and measurement-intensive system should be used as a continuing performance- and productivity-improvement system.

If we plan from the first to use evaluation data continually to improve performance and productivity, the need for summative evaluation is eliminated. Performance on reference tasks can be noted all along the way, and when performance is anchored to a measurement scale, it will generally provide more interesting and understandable information than score gains. Summative evaluation in training is at best a snapshot at a single point in time of one attempt to improve performance. The concepts of continuous measurement discussed in this paper, and embodied in the concept of mastery assessment systems, make it possible to observe progress in groups or individuals at many points in time. We need no pre-set standard to judge our success, but can continue to achieve performance improvements through the use of continuous measurement. For example, if the management group in the scenario had set for themselves a target of, say, 90% of the exemplary performance identified in the highest performers, this very target should later give way as improvements in technology and in the proficiency of the work force enable workers to set higher targets. Certainly such a continually improving system is an important response to the great social need for global economic competitiveness. The untapped potential for productivity improvement is there. Alan Greenspan (1988), quoted earlier, sees cause for hope:

> "...the potential for further gains in efficiency is immense. In part, it will require building on those adjustments that already have been put in place. But we also must take a broader view and

begin to make adjustments that will allow more scope in the work place for individual initiative and enterprise. And it is time to consider alternative approaches to management and the organization of work." (p. 10)

Conclusions

This paper was designed to introduce training practitioners to some new concepts about how measurement science can provide a new framework for assessing progress and can add new discipline to the development, implementation, and conduct of training. Revitalized by interactive technologies and the disciplines of cognitive science, instructional science, applied artificial intelligence, and the studies of organizational change, measurement science could be used to address the economic challenge our nation currently faces--the demand for a much smarter, much more productive work force, able to deal with high levels of technology and automation flexibly and productively. High levels of expertise could be attained by the nation's work force through the application of considerably deeper and more disciplined uses of science and technology and their handmaiden, measurement, in facilitating the attainment of the desired expertise.

Deeper and more multi-disciplinary methods of front-end analysis will be necessary. We need accurate models of the domains of knowledge to be traversed. We need human exemplars of high competence. We need maps of the domain to show the journey to the trainees and to track progress. We need reference tasks that allow performance to be measured at each of the stages of growth, from novice through intermediate to expert. We need cognitive and psychometric analyses to show the range of scales that span the domain. We need to understand the range of constructs, especially those that characterize expertise along the path, and to include deep analyses of constructs at the points of discontinuity where knowledge structures change in going from one level of expertise to another.

It was argued that we need new methods of instructional development that are more iterative and that benefit from formative evaluation. Such methods of development must be tied closely to implementation plans and strategies. They must be designed to produce organizational change through an incremental strategy rather than through the rapid and usually unsuccessful introduction of massive technological change in a resistant organization. An ongoing formative evaluation should produce incremental improvement, not only in the instructional materials, but in the implementation plans and tactics as well.

This paper has deliberately projected a speculative and complex scenario requiring a mixture of contributions from several different fields: training and training systems development, performance engineering, interactive systems applied to teaching and testing, knowledge engineering, allied cognitive science, instructional science, and educational measurement. The goal is to find a new mixture that will bring about the possibility of massive improvements in performance and productivity through better use of technology, training, and implementation. Reaching this goal is not the task of measurement science alone, nor of training practice alone. It requires collaboration and cross-fertilization among disciplines in responding to the high challenge our nation faces. It is hoped that this paper will generate a dialogue that will motivate greater contributions from each discipline and provide a basis for future combined efforts.

References

Bailey, R. W. (1982). Human performance engineering: A guide for system designers. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1977). Social learning theory. Englewood Cliffs, NJ: Prentice-Hall.

Bennett, R.E., Gong, B., Kershaw, R.C., Rock, D.A., Soloway, E., & Macalalad, A. (1988). Agreement between expert system and human ratings of constructed responses to computer science problems (Report No. RR-88-20). Princeton, N.J: Educational Testing Service.

Branson, R.J. & Grow, G. (1987). Instructional systems development. In R. N. Gagne (Ed.), Instructional technology: Foundations. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bruner, J.S. (1968). Toward a theory of instruction. New York: W.W. Norton.

Bunderson, C.V. (1987). The modelling of expertise. In Proceedings of the Department of Defense/Educational Testing Service Conference on Job Performance Measurement Technologies. Washington, DC: Office of the Assistant Secretary of Defense.

Bunderson, C.V., Inouye, D.I., and Olsen, J.B. (in press). The four generations of computerized educational measurement. In R. Linn (Ed.), Educational measurement (3rd ed.). New York: Macmillan.

Carnegie Forum on Education and the Economy. (1986). A nation prepared: Teachers for the twenty-first century. Washington, DC: Author.

Fine, S.A., & Wiley, W. (1971). An introduction to functional Job analysis. Methods for manpower analysis #4. Washington, DC: W. E. Upjohn Institute for Employment Research.

Forehand, G.A., & Bunderson, C.V. (1987). Basic concepts of mastery assessment systems. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Forehand, G.A., & Bunderson, C.V. (1987). Mastery assessment systems and educational objectives. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Gilbert, F. (1978). Human competence: Engineering worthy performance. New York: McGraw-Hill.

Greenspan, A. (1988, January). Remarks. Paper presented at the Martin Luther King, Jr. Social Responsibility Seminar, Atlanta, GA.

Hart, A. (1986). Knowledge acquisition for expert systems. New York: McGraw-Hill.

Johnson, W.L., & Soloway, E. (1987). PROUST: An automatic debugger for Pascal programs. In G. Kearsley (Ed.), Artificial intelligence and instruction: Applications and methods. Reading, Mass: Addison-Wesley.

Kirsch, I.S. (1987, February). Measuring adult literacy. Toward defining literacy. Symposium sponsored by National Advisory Council on Adult Education, Literacy Research Center, University of Pennsylvania, Philadelphia, PA.

Kirsch, I.S., & Jungeblut, A. (1986). Literacy: Profiles of America's young adults--Final report (NAEP Report No. 16-PL-01). Princeton, NJ: National Assessment of Educational Progress.

Lipson, J.I. (1988). Testing in the service of learning science--assessment systems that promote educational excellence and equality. Assessment in the Service of Learning: Proceedi.ıgs of the 1987 ETS Invitational Conference. Princeton, NJ: ETS.

McCormick, E.J. (1979). Job analysis; Methods and applications. New York: Amacon, a division of American Management Associations.

Naisbitt, J. (1984). Megatrends. New York: Warner Books.

Polanyi, M. (1962). Personal knowledge: Towards a post-critical philosophy. New York: Harper & Row.

Simon, H.A. (1981). The sciences of the artificial. Cambridge, MA: MIT Press.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 195-201.

Author Notes

I am indebted to my colleagues at Educational Testing Service--Norman Frederiksen, Garlie Forehand, Irwin Kirsch, Myrtle Rice, Mike Rosenfeld, and Ben Shimberg--for providing valuable reviews and editorial contributions from the perspectives of measurement science, cognitive science, industrial psychology, and writing style; also to Deane Gradous and Cathy Sleezer for their thorough editing. Any remaining mistakes or omissions are my own.

Footnotes

[1] This material is adapted and updated from discussions of mastery systems first described by Forehand and Bunderson (1987a, 1987b), and similar material found in Bunderson, Inouye, Olsen (in press).

Table 1

Test Items vs. Reference Tasks

Test Items	Reference Tasks
1) Usually administered via paper and pencil.	1) Usually administered via interactive computer.
2) Written objectives prescribe test items.	2) Flow-charts/interaction specifications prescribe reference tasks.
3) Single response, usually multiple choice	3) Multiple responses, which together provide for a quantitative assessment of degree of success.
4) Dichotomous scoring o Pass o Fail	4) Trichotomous scoring o Pass (competence demonstrated) o Needs coaching and practice o Not ready for this task
5) A complete test with subtests can be used for diagnostic purposes.	5) Simultaneous subscores are taken to measure component processes/states. These provide data to guide coaching.

(table continues)

Test Items	Reference Tasks
6) Items and entire tests are often decontextualized. Lay people may not see the relevance of the questions to valued capabilities in the real world.	6) Tasks refer to or simulate aspects of valued real-world activity (e.g., in a job).
7) Items can be calibrated and placed on a measurement scale.	7) Reference tasks are related to one or more statistically validated scales which span the domain. The calibration of reference tasks is a current challenge for measurement science.
8) Except in CAT systems, administration of the next item is fixed by its order on the page.	8) Next response request is determined dynamically.
9) Test items are "used up" after one attempt and are of little value for repeated practice.	9) While some reference tasks (e.g., paragraphs to read) require files of alternative stimuli for practice, many of them (e.g. simulations or game-like events) can be practiced repeatedly without using up material.

(table continues)

Test Items	Reference Tasks
10) The objectives and specifications used to develop test items are not suitable for learners, and they are not presented.	10) A "model of mastery" can be made available to help learners observe the attributes of successful performance.