ED 376 180 TM 022 168

AUTHOR Mislevy, Robert J.
TITLE Test Theory Reconceiv d. Project 2.4 Quantitative Models To Monitor th Status and Progress of Learning and Performance and Their Antecedents.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Educational Testing Service, Princeton, N.J.; Office of Educational Research and Improvement (ED), Washington, DC.; Office of Naval Research, Washington, D.C.
PUB DATE May 94
CONTRACT N00014-91-J-4101; R117G10027
NOTE 66p.; Version of a paper presented at the Annual Meeting of the National Council of M surement in Education (Atlanta, GA, April 13-1 '93).
PUB TYPE Reports - Evaluative/Feasibility (Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Cognitive Psychology; *Educational Assessment; *Inferences; Measurement Techniques; Models; Psychometrics; Statistical Analysis; Test Construction; Testing; *Test Theory
IDENTIFIERS *Alternative Assessment

ABSTRACT

Test theory encompasses models and methods for drawing inferences about what students know and can do, cast in a framework of ideas from measurement, education, and psychology. The emerging paradigm of cognitive psychology prompts new considerations about collecting and interpreting evidence, suggesting alternative models for the nature, acquisition, and assessment of competence. Aspects of the models and methods that have been developed in the framework of standard test theory can be extended to the new discourse about student learning, but it is necessary to disentangle statistics from psychology, and to distinguish how we are reasoning from what we are reasoning about. Toward this end, the interplay of reasoning per se and the universe of discourse in which a problem is framed are discussed. Educational testing within alternative psychological paradigms and the inferential tasks entailed are considered. Implications of cognitive psychology for test theory are discussed and illustrated with examples from current projects. Five figures and six tables illustrate the discussion. (Contains 90 references.) (SLD)

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

Final Deliverable – May 1994

Project 2.4 Quantitative Models to Monitor the Status and Progress of Learning and Performance and Their Antecedents

Test Theory Reconceived

Robert J. Mislevy, Project Director
CRESST/Educational Testing Service

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027   CFDA Catalog No. 84.117G

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

The work reported herein was supported in part by the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education; Contract No. N00014-91-J-4101, R&T 4421573-01, from the Cognitive Science Program, Cognitive and Neural Sciences Division, Office of Naval Research; and the Statistical and Psychometric Research Division of Educational Testing Service.

The findings ar·1 opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

To this end, the following section discusses the interplay of reasoning *per se* and the universe of discourse in which a problem is framed. Educational testing within alternative psychological paradigms, and inferential tasks thus entailed, are then considered. Implications of cognitive psychology for test theory are discussed and illustrated with examples from current projects.

Evidence and Inference

Inference is reasoning from what we know and what we observe to explanations, conclusions, or predictions. We are always reasoning in the presence of uncertainty. The information we work with is typically incomplete, inconclusive, amenable to more than one explanation. We must apply in educational assessment many of the same skills needed in such fields as troubleshooting, medical diagnosis, and intelligence analysis. We attempt to establish the weight and coverage of evidence in what we observe. But the very first question we must address is "Evidence about what?" There is a crucial distinction between *data* and *evidence*: "A datum becomes evidence in some analytic problem when its *relevance* to one or more hypotheses being considered is established. . . . [E]vidence is relevant on some hypothesis if it either increases or decreases the likeliness of the hypothesis. Without hypotheses, the relevance of no datum could be established" (Schum, 1987, p. 16).

Test data, like clues in a criminal investigation, acquire meaning only in relation to a network of conjectures. The same observation can be direct evidence for some conjectures and indirect evidence for others, and wholly irrelevant to still others. In criminal investigations, we construct our conjectures around notions of the nature of crime, of justice, of proof, of human nature itself (compare the proceedings of contemporary trials with those of the Inquisition). The conjecture we might entertain under one conception of justice, let alone the kind of data we would seek to support it, might not even be possible to express under an alternative conception. In educational assessment, we construct our conjectures around notions of the nature and acquisition of knowledge and skill.

An example hints at directions we need to explore. The Mathematical Sciences Education Board (MSEB) recently published a collection of prototype assessment tasks designed to allow children to "demonstrate the full range of

TEST THEORY RECONCEIVED[1]

Robert J. Mislevy
Educational Testing Service/CRESST

Introduction

Test theory, as we usually think of it, is part of a package. It encompasses models and methods for drawing inferences about what students know and can do—as cast in a particular framework of ideas from measurement, education, and psychology. This framework generates a universe of discourse: the nature of the problems one defines, the kinds of statements one makes about students, the ways one gathers data to support them. Test theory, as we usually think of it, is machinery for inference within this framework.

The emerging paradigm of cognitive psychology also generates a universe of discourse, engendering its own kinds of scientific and applied problems, suggesting alternative models for the nature and the acquisition of competence, prompting new considerations about how to collect and interpret evidence. Just as under the standard testing paradigm, however, we face such questions as: What kinds of evidence are needed to support inferences about students? How much faith can we place in the evidence, and in the statements? Are elements of evidence overlapping, redundant, or contradictory? When must we ask different questions or pose additional situations to distinguish among competing explanations of what we see? Aspects of the models and methods that have been developed within the framework of standard test theory can be extended, augmented, and reconceived to address problems cast in a broader universe of discourse about students' learning. It is necessary, however, to disentangle the statistics from the psychology in standard test theory; to distinguish *how we are reasoning* from *what we are reasoning about*.

[1] An earlier version of this paper was presented at the annual meeting of the National Council of Measurement in Education, Atlanta, April 1993. Comments by the discussants Bob Glaser, H.D. Hoover, and Dick Snow have been incorporated, along with comments from Isaac Bejar, Kalle Gerritz, and Howard Wainer.

their mathematical power, including such important facets as communication, problem-solving, inventiveness, persistence, and curiosity" (MSEB, 1993, p. iii). Figure 1 is part of one task. A National Research Council newsletter stated that "rather than focusing on rote recall and routine arithmetic, [the prototypes] measure the ability to understand and apply higher-level concepts" (Push & Hicks, 1993, p. 7).

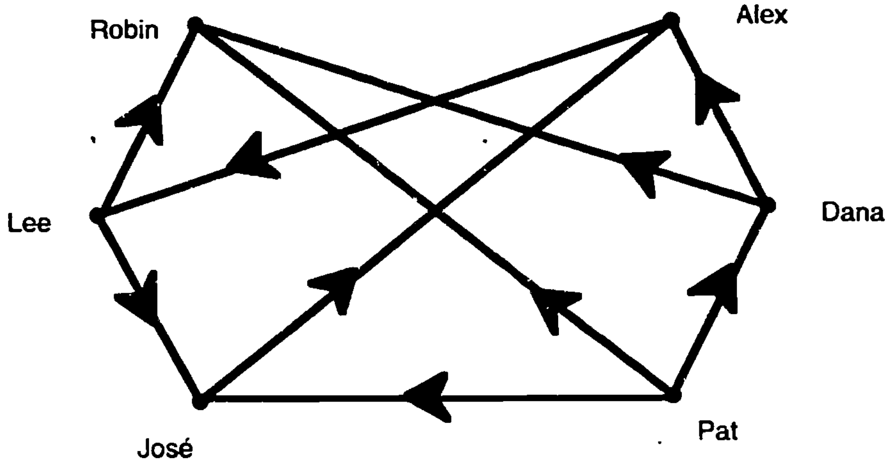
A graph and a series of questions may indeed stimulate interesting mathematical thinking on the part of students. They may, further, evoke behavior that tells us something about that thinking—data that may turn out to be useful evidence for conjectures we are interested in. But they do not, in and of themselves, "measure" anything. Exactly what aspects of thinking do we want to talk about, and how do we relate what we observe in this specific situation to a more abstract level of discourse? Do we want to speak beyond this particular graph and set of questions, to, say, how students might handle different questions about the same graph? Or similar graphs with different questions? Or tasks that don't involve graphs at all, but require explanations of mathematical concepts? Should we summarize our observations in terms of a single aspect of students' solutions or many? In terms of numbers, ordered categories, qualitative distinctions, or some mixture of these? We must start by determining just what we want to talk about.

Paradigms

Thomas Kuhn (1970) used the term "paradigm" to describe a set of interrelated concepts that frames research in a scientific field. A paradigm gives rise to what I've been calling a "universe of discourse." Of all the phenomena that we can experience directly or indirectly, a paradigm focuses on patterns in a circumscribed domain. The patterns determine the kinds of things we talk about; the characteristics, the particular things we say. (In formal scientific work, the patterns might be expressed as models; the characteristics, as values of variables in models.) A paradigm determines what we construe as problems, and how we evaluate our attempts to solve them. Some examples of paradigms are Newtonian and quantum mechanics; the geocentric and the heliocentric views of solar system; and, most pertinent to our present concerns, trait, behavioral, and cognitive psychological paradigms.

Prototype from *Measuring Up*

Six children are in a checkers tournament. The figure below shows the results of the games played so far. Arrows point in the direction of the loser. For example, Alex won his game against Lee.



1. Who won the game between Pat and Robin?
2. Make a table showing the current standings of the six children.
3. Dana and Lee have not played yet. Who do you think will win when they play? Explain why you think so.

Figure 1. A mathematics task prototype.

No paradigm is all-encompassing. Birnbaum (1991, p. 65) describes the view that ". . . problem-solving depends on the manipulation of relatively fragmented and mutually inconsistent *microtheories*—each perhaps internally consistent, and each constituting a valid way of looking at a problem: 'This will allow us to say, for example, that some [set of beliefs] is more appropriate than some [other set of beliefs] when confronted with problems of diagnosing bacterial infections. Scientists are used to having different—even contradictory—theories to explain reality . . . Each is useful in certain circumstances' (Nilsson, 1991, p. 45)."

Sometimes paradigms address overlapping phenomena. When two observers view the same event through the lens of different paradigms, however, they attend to different aspects of what they see, and make different connections to other concepts. Where Priestly "saw" dephlogistated air, Lavoisier "saw" oxygen (Kuhn, 1970, p. 118). Confusion reigns when different paradigms use the same words with different meanings, as the same observation can lead to contradictory conclusions. We shall discuss an example from test theory below, concerning how to "account for the difficulty" of assessment tasks.

Most scientific research is carried out within an existing paradigm. Kuhn used the term "normal science" for solving the outstanding puzzles a paradigm poses. Normal science improves measurements, develops inferential machinery, works out relationships in greater detail, extends ideas to new situations, and integrates previously separate elements. Applied problem solving takes the same flavor. The concepts and patterns of a paradigm are taken as givens, into which the elements of a particular application are mapped. These structures guide data gathering, interpretation, and decision making.

Kuhn studied "scientific revolutions," in which a new major paradigm displaces an existing paradigm. A paradigm shift can be precipitated by a paradigm's failure to deal with some outstanding problem—perhaps a puzzle that is intractable as framed in the existing paradigm, or a problem it cannot frame at all. New concepts arise; new relationships are highlighted. Some concepts and relationships overlap with those of the previous paradigm, as do methodologies and phenomena addressed, but the essential organizing structure changes. A paradigm shift redefines what scientists see as problems, and reconstitutes their tool kit for solving them. Previous models and methods remain useful to the extent that certain problems the old paradigm addresses are still meaningful, and the solutions it offers are still satisfactory, but now as viewed from the perspective of the new paradigm.

Psychological Paradigms and Test Theory

> Particular forms of tests and assessments represent particular forms of discourse, that is, they produce particular ways of talking and communicating with others about the schooling and education process. (Berlak, 1992, p. 186)

I like this quotation for two reasons. The first is that it connects how we think about assessing with how we think about learning and teaching. The second was actually a reason I *didn't* like it when I first saw it: the order is backwards. A conception of student competence and a purpose for assessment should determine the kind of information one needs, which should in turn suggest ways to get students to reveal something about their competencies, that is, the forms of assessment. But Berlak's description does reflect common practice. Too often, an assessment form is adopted without conscious consideration of the purpose of the assessment and the nature of competence that should underlie the effort. A universe of discourse is instantiated by default, often presuming concepts and values of the paradigm that gave rise to that form of assessment.

Making a rational choice of assessment methods requires thinking these issues through. The following sections discuss implications that the trait, behaviorist, and cognitive psychological paradigms hold for test theory. We cannot deeply pursue here all the ways in which different purposes entail different evidential requirements, even under a given conception of competence (see Millman & Greene, 1989, on dimensions of purpose that also shape the form of assessment). Purposes mentioned in the following discussion include selecting students into fixed alternatives, monitoring the progress of groups of students, and planning instruction for individual students.

Trait Psychology and "Mental Measurement"

The most familiar tools of standard test theory began to evolve a century ago under the paradigm of trait psychology, initially in a quest to "measure people's intelligence." Messick (1989, p. 15) defines a trait as "a relatively stable characteristic of a person—an attribute, enduring process, or disposition—which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances." These invented (hence, inherently unobservable) numbers are proposed to locate people along continua of mental characteristics, just as their heights and weights locate them along continua of physical characteristics.

When Spearman used scores on knowledge and puzzle-solving tasks to "measure intelligence," the notion of a trait was not new. Paul Broca and Francis Galton had attempted to assess "intelligence" in the previous century, Broca by charting cranial volumes, Galton by measuring reaction times. Nor was the idea of observing behavior in samples of standardized situations new. Three thousand years ago the Chinese discovered that observation of an individual's performance under controlled conditions could support accurate predictions of performance under broader conditions over a longer period of time (Wainer et al., 1990, p. 2). The essence of mental measurement was, rather, a confluence of these concepts: identifying "traits" with tendencies to behave in prescribed ways in these prescribed situations. Variables so defined were viewed as *the* way to characterize people—the psychology—and test scores as *the* way to obtain the requisite evidence—the methodology: "Intelligence is what tests of intelligence test, until further scientific observation allows us to extend the definition" (Boring, 1923, p. 35). As in physical measurement, great care was taken to define the tasks, the conditions under which they were administered, and the rules for mapping observations to summary scores.

This psychology and the methodology suited the mass educational system that also arose in the United States at the turn of the century (Glaser, 1981). Educators viewed their challenge as selecting or placing large numbers of students in instructional programs, when resources limited the amount of information they could gather about each student, constrained the number of options they could offer, and precluded much tailoring of programs to individual students once the decision was made. This view of the problem context encouraged building student models around characteristics that were few in numbe:, broadly construed, stable over time, applicable to wide ranges of students, and discernible with data that were easy to gather and interpret.

A Brief History of Test Theory

Test theory research over the century exhibits the extensions, generalizations, and increasing technical sophistication *within a given paradigm* that mark "normal science"—in this case, within the paradigm of characterizing people's tendencies to behave in prescribed ways in prescribed settings. The inferential considerations that motivate these developments merit a brief review because they transcend the substantive content of the psychological paradigm under which test theory arose; analogous considerations arise no matter which psychological paradigm underlies an assessment. We highlight the interplay between the substantive content of the paradigm (the semantics) and the methodology of reasoning within the paradigm (the syntax).

Edgeworth (1888, 1892) and Spearman (1904, 1907) launched classical test theory (CTT) by applying mathematical models and statistical tools from physical measurement to what were, under the paradigm, comparable problems in mental measurement. CTT views the average of 1-for-right/0-for-wrong results from a sample of test items from a domain as a noisy measure of an examinee's "true score." While each individual item taps different skills and knowledge in different ways for different people, a total score accumulates over items a general tendency to answer items from the domain correctly, and conveys direct evidence for conjectures about a variable so construed (Green, 1978). Different random samples of tasks from the same domain, or parallel tests, are alternate sources of information about tendencies to behave in the prescribed manner in these situations. Scores on parallel tests are direct evidence, each with the same amount of weight and the same scope of coverage, about the same true score.

The key inferential concept in test theory is *conditional independence*. Stated generally, variables may be related in a population, but independent given the values of another set of variables. The paradigm of a field supplies concepts, variables, and conditional independence relationships. In CTT, interest centers on the unobservable variable "true score," with observable scores on parallel tests posited to be conditionally independent given true score. Judah Pearl argues that inventing intervening variables such as true scores is not merely a technical convenience, but a natural element in human reasoning:

> [C]onditional independence is not a grace of nature for which we must wait passively, but rather a psychological necessity which we satisfy actively by organizing our knowledge in a specific way. An important tool in such organization is the identification of intermediate variables that induce conditional independence among observables; if such variables are not in our vocabulary, we create them. In medical diagnosis, for instance, when some symptoms directly influence one another, the medical profession invents a name for that interaction (e.g., "syndrome," "complication," "pathological state") and treats it as a new auxiliary variable that induces conditional independence; dependency between any two interacting systems is fully attributed to the dependencies of each on the auxiliary variable. (Pearl, 1988, p. 44)

Spearman's *methodological* insight (as distinct from his thoughts about human abilities *per se*) was this: conditional independence of observable test scores, given an unobservable "intelligence" variable, would imply particular patterns of relationships among the observable scores (Spearman, 1904, 1927). This insight provides a framework for organizing observations, and for quantifying and (at least in principle) disconfirming conjectures about behavior in terms of hypothesized traits. Test theorists have since been working out the logic of inference in terms of unobservable variables: exploring the possibilities and the limitations, developing statistical machinery for estimation and prediction—in short, learning how to reason within the paradigm of mental measurement.

The original indicator of a test's evidential value under CTT was *reliability*, the correlation between parallel forms in a specified population of examinees.[2] This definition reflects the classic norm-referenced usage of tests: locating people along a single dimension, for selection and placement decisions. A high reliability coefficient indicates that a different sample of tasks of the same kind would order the examinees similarly, leading to the same decision about most of them. Reliability is a sensible summary of the

[2] Even if only one form of a test existed, an estimate of its reliability could be obtained nevertheless from the internal consistency of its constituent elements; e.g., for tests of exchangeable items, the average correlation among all possible half-tests, adjusted upwards to account for their shorter length.

evidence a test provides *in this specific context* (a particular group of students and a domain of tasks), *for this specific purpose* (lining the students up comparatively for selection or placement) *under this specific psychological paradigm* (assuming that lining them up according to true scores would capture what matters). Reliability does not characterize the evidence test scores might provide about other conjectures, even those framed within the CTT paradigm; for example, whether a student's true score is above a specified cutoff value, or the magnitude of change in true score from pretest to posttest.

Extending the Methodology to Behavioral Psychology

Messick's phrase "relatively stable" softens the extreme early conception of a trait—which might be described as "inborn and unchangeable"—and acknowledges the extended range of phenomena to which the models and methods of CTT came to be applied. We hope that a student's tendency to perform well on mathematics tasks *will* change, through instruction and experience. At any given point in time, however, one might contemplate gauging her overall proficiency with respect to specified domains of tasks, as defined perhaps by this week's lesson, or by a consensually defined collection that "a minimally competent eighth grader" in her state "should be able to answer." This usage extends the application of CTT machinery beyond the original selection and placement decisions, to planning and evaluating instruction from the perspective of behavioral psychology:

> The educational process consists of providing a series of environments that permit the student to learn new behaviors or modify or eliminate existing behaviors and to practice these behaviors to the point that he displays them at some reasonably satisfactory level of competence and regularity under appropriate circumstances. The statement of objectives becomes the description of behaviors that the student is expected to display with some regularity. The evaluation of the success of instruction and of the student's learning becomes a matter of placing the student in a sample of situations in which the different learned behaviors may appropriately occur and noting the frequency and accuracy with which they do occur. (Krathwohl & Payne, 1971, pp. 17-18)

The familiar standardized achievement test consists of a sample of tasks in an area of learning, and students' "true scores" are tendencies to make correct responses rather than incorrect responses, for example, or to write coherent rather than disjointed essays. The object of inference in this case is not a "trait" in Galton's or Spearman's sense, but simply a summary of a behavioral tendency in a class of stimulus situations—an "overall proficiency" in the prescribed domain of tasks. CTT's data-gathering methodologies and inferential machinery for summarizing behavior in samples of prescribed situations were thus extended to instructional problems cast in behavioral psychology.

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) broadened the notion of the evidential value of an observed test score, taking into account the conditions under which the data were obtained and how they were to be used. The statistical machinery of generalizability theory first characterizes the variation associated with facets of observation, such as samples of tasks and students, and, when judgment is involved, numbers and assignment patterns of raters. It can then quantify the evidence that scores from a observational setting convey for such various inferences as comparisons between examinees, of examinees against a fixed criterion, and of changes over time; in terms of the domain of tasks as whole, with different numbers or kinds of raters, in different subdomains (e.g., what does a student's narrative essay tell us about how well she can write friendly letters?), and so on. Generalizability theory expands the range of conjectures one can address, but still within a universe of discourse in which inferences still concern "overall tendency toward specified behavior in a specified domain," as defined from the point of view of the test designer.

Item Response Theory

A source of dissatisfaction with CTT early on was that its characterizations of examinees (e.g., domain true score and percentile rank) and tasks (e.g., percent-correct and item-test correlation) were tied to specific collections of examinees and tasks. Item response theory (IRT; see Hambleton, 1989, for an overview) originated in the early 1940s as an attempt to characterize examinees' proficiency independently of the tasks they happened to have taken, and tasks independently of the examinees who happened to take them—a goal inspired by the analogy to physical measurement. Like CTT, IRT addresses examinees' proficiency in a domain of tasks. Beyond CTT, IRT posits a functional relationship between proficiency and probability of correct response to a given item that is the same for all examinees. That is, differences in students' chances of success are modeled as depending solely on their values on the single overall-proficiency variable (e.g., tendency to mark correct answers in this domain of test items). Such a pattern cannot be expected a priori for simply any domain of tasks and any group of students. (And, as we discuss below, there are inferences for which this pattern is not the one we need to model!) However, when we do successfully construct an assessment context in which observations do approximate this pattern, we are justified in using a formal measurement framework to guide inference in that context (Wright, 1977).

IRT helped solve practical problems that could be expressed in the mental measurement paradigm but were poorly handled with CTT tools, such as constructing tests with desired properties and tailoring tests to individual examinees.[3] The IRT formulation lends itself well to the machinery of statistical inference. The relationships among observable variables, and by implication between observable and hypothesized unobservable variables, are laid out more explicitly than in CTT. Rapid progress has been made by applying recent developments in statistics to IRT (e.g., Bock & Aitkin, 1981; Lord, 1980; Mislevy, 1991). And Georg Rasch (1960) solved a central theoretical question of mental measurement by explicating the class of models under which, if true, examinees could be compared independently of the items they responded to, and items compared regardless of the sample of examinees who responded to them.[4] Note that these are all issues of how to reason within a paradigm, of syntax within a universe of discourse. Determining the real-world contexts for which the models are appropriate is quite a separate issue.

In statistical framework, estimation tools strengthen inference under the assumption that a model is correct. Just as importantly, however, diagnostic tools help determine when and where the model fails—at once improving

[3] This is due to the use of a more powerful conditional independence relationship. Rather than CTT's *test level* conditional independence of scores on parallel tests given true score, IRT is based on *item level* conditional independence, namely responses to items given the hypothetical proficiency variable. One can thus combine evidence from individual test items in far more flexible arrangements than parallel tests—at the cost, of course, of verifying a more restrictive model. It is important in a given application to explore how the particular ways the model fails to fit will affect the particular inferences one wants to make.

[4] See Andrich (1988) for a discussion of Rasch's approach as a paradigm shift in test theory.

applications within the paradigm and providing clues to see beyond it: "To the extent that measurement and quantitative technique play an especially significant role in scientific discovery, they do so precisely because, by displaying serious anomaly, they tell scientists when and where to look for new qualitative phenomena" (Kuhn, 1970, p. 205). We shall say more in Example 1 about the importance of diagnostic tools for using IRT in light of results from cognitive psychology.

In addition to IRT, a separate stream of test theory research has been the analysis of relationships among scores from different tests. Factor analysis, structural equations modeling, and multitrait-multimethod analysis all address patterns in correlations among scores of several tests, in hopes of better understanding the meaning of variables so defined. A researcher might seek to identify recurring patterns in tests with systematically varying tasks; for example, looking for broadly defined tendency to perform well on scientific inquiry tasks, using scores from multiple-choice items, computer simulations, and laboratory notebooks (Shavelson, Baxter, & Pine, 1992). Additional tests with the same formats, but with, say, mathematics content, might be added to see whether examinees vary systematically as to their performance in various formats, as distinct from their proficiencies in the content areas (Campbell & Fiske, 1959).

These correlational tools are the main way test theorists have sought to establish the weight and coverage of evidence test scores provide for inferences—in a word, validity. Early selection and placement applications focused exclusively on the correlation between the scores used to make decisions and the scores summarizing outcomes of subsequent programs, calling this number *the* validity coefficient. Contemporary views of validity even within the paradigm (Messick, 1989) are considerably broader:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. . . . [W]hat is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators—inferences about score meaning or interpretation and about the implications for action that the interpretation entails . . .
>
> Different kinds of inferences from test scores may require a different balancing of evidence, that is, different relative emphases in the range of evidence presented.

> By *evidence* is meant both data, or facts, and the rationale or arguments that cement those facts into a justification of test-score inferences. . . . One or another . . . forms of evidence, or combinations thereof, have in the past been accorded special status as a so-called "type of validity" [e.g., content, criterion, predictive, concurrent, and construct validity]. But because all of these forms of evidence bear fundamentally on the valid interpretation and use of scores, it is not a type of validity but the relation between the evidence and the inferences drawn that should determine the validation focus. (Messick, 1989, p. 13ff)

At its leading edge, if not in everyday practice, test theory for the mental measurement paradigm has come of age—in the sense of having developed methodological tools for gathering and interpreting data, and a coherent conceptual framework for inference about students' tendencies to prescribed behaviors in prescribed settings. *The question is the extent to which the inferences we now want to make for guiding and evaluating education can be framed within this universe of discourse.*

What Overall-Proficiency Measures Miss

Evidence can now be brought to bear on inferences about students' overall proficiency in behavioral domains, for determining a student's level of proficiency, comparing him to others or to a standard, or gauging change from one point in time to another. Summarizing competence in these terms suits the kinds of low-resource, long-lasting decisions it was designed for: sorting, assigning, or selecting students into educational activities—presumably with the general objective of helping students become more proficient. Conjectures about the nature of this proficiency or how it develops fall largely outside the mental-measurement paradigm's universe of discourse. As Stake (1991, p. 245) notes, "The teacher sees education in terms of mastery of specific knowledge and sophistication in the performance of specific tasks, not in terms of literacy or the many psychological traits commonly defined by our tests."

Cronbach and Furby's (1970) "How should we measure 'change'—or should we?" reflects the frustration of recognizing vital questions beyond a paradigm's reach. After cogently analyzing the subtleties of inference about change under CTT, they lament an overarching inadequacy of all of the techniques they discuss:

> Even when [test scores] X and Y are determined by the same operation [e.g., a given CTT or IRT model for specified behavior in a specified domain of tasks], they often do not represent the same psychological processes (Lord, 1958). At different stages of practice or development different processes contribute to the performance of a task. Nor is this merely a matter of increased complexity; some processes drop out, some remain but contribute nothing to individual differences within an age group, some are replaced by qualitatively different processes. (p. 76)

The criterion-referenced testing movement of the 1960s (e.g., Glaser, 1963) attempted to bring the machinery of the mental measurement paradigm to bear on instructional problems by defining behavioral domains with greater specificity, so that educators could infer in detail what students could and could not do. Merely providing detailed descriptions of performance proves insufficient to make test scores relevant, however, if it fails to address the underlying knowledge, skills, and strategies that lead to performance and serve as the foundation for further development (Glaser, 1981).

Implications of Cognitive Psycholc gy for Test Theory

Most contemporary research into human abilities takes place within neither the trait nor behavioral psychological paradigms, but within what has come to be called the cognitive paradigm. Cognitive functions include "such activities as perceiving relationships, comparing and judging similarities and differences, coding information into progressively more abstract forms, classification and categorization, memory search and retrieval" (Estes, 1981, p. 11), and, more to our point, learning and problem solving. Cognitive psychology explores just how it is that people do these things. Three working propositions from cognitive psychology (paraphrasing Lesh & Lamon, 1992, p. 60) hold implications for education:

1. People interpret experience and solve problems by mapping them to internal models.

2. These internal models must be constructed.

3. Constructed models result in situated knowledge that is gradually extended and decontextualized to interpret other structurally similar situations. With use, aspects of mapping and problem solving become automated.

People Interpret Experiences and Solve Problems by Mapping Them to Models

Knowledge structures have been studied as "mental models" (Johnson-Laird, 1983), "frames" (Minsky, 1975), and "schemas" (Rumelhart, 1980). A schema (using Rumelhart's term inclusively for convenience) can be roughly thought of as a pattern of recurring relationships, with variables that in part determine its range of applicability. Associated with this knowledge are conditions for its use. Rumelhart (1980, p. 33ff) views schemas as *"the building blocks"* of cognition: "Schemata are employed in the process of interpreting sensory data (both linguistic and nonlinguistic), in retrieving information from memory, in organizing actions, in determining goals and subgoals, in allocating resources, and, generally, in guiding the flow of processing in the system."

Moreover, "it looks like schemas are the key to understanding expertise" (VanLehn, 1988, p. 49). While experts in various fields of learning do generally command more facts and concepts than novices, and have richer interconnections among them, a key distinction lies in their ways of viewing phenomena, and representing and approaching problems (e.g., Chi, Feltovich, & Glaser, 1981, on physics; Lesgold, Feltovich, Glaser, & Wang, 1981, on radiology; and Vcss, Greene, Post, & Penner, 1983, on social science). The advanced concepts that college physics students acquire can be organized around informal associations or naive misconceptions (Caramazza, McCloskey, & Green, 1981). They tackle problems less effectively than expert physicists, whose more appropriate schemas lead them to the crux of the matter (Chi et al., 1981):

> Schemata play a central role in all our reasoning processes. Most of the reasoning we do apparently does not involve the application of general purpose reasoning skills. Rather, it seems that most of our reasoning ability is tied to particular bodies of knowledge. . . . Once we can "understand" the situation by encoding it in terms of a relatively rich set of schemata, the conceptual constraints of the schemata can be brought into play and the problem readily solved. It is as if the schema already contains all of the reasoning mechanism ordinarily required in the use of the schemata. Thus, understanding the problem and solving it is nearly the same thing. (Rumelhar+, 1980, p. 55ff)

Internal Models Must Be Constructed

A schema is "instantiated" when we perceive some of its relationships in a situation, which focuses our attention on filling in other variables, inferring additional relationships, and checking for specifics at odds with usual expectations. Much of this activity is unconscious and automatic, as when we process individual letters in the course of reading a text. Sometimes aspects of it are conscious and deliberate, as when we try to determine the text's implications. "The total set of schemata instantiated at a particular moment in time constitutes our internal model of the situation we face at that moment in time" (Rumelhart, 1980, p. 37). No act of cognition is purely passive or data-driven; we must ever and always construct meaning, in terms of knowledge structures we have created up to that point in time. Thus,

> . . . it is useful to think of a schema as a kind of informal, private, unarticulated theory about the nature of events, objects, or situations that we face. The total set of schemata we have available for interpreting our world constitutes our private theory of the nature of reality. (Rumelhart, 1980, p. 37)

Situated Knowledge Is Extended and Decontextualized; Procedures Are Automated.

If perception is an active process (selecting, building, and tailoring representations from currently available schemas), then learning is all the more dynamic: extending, modifying, and replacing elements to create new structures. In some cases learning is merely adding bits to existing structures. Sometimes it involves generalizing or connecting schemas. Other times it involves wholesale abandonment of major parts of schemas, with replacement by qualitatively different structures (Rumelhart, 1980). The parallels between the development of personal knowledge within an individual and public knowledge in scientific community have not gone unnoticed:

> The process of knowledge acquisition can be conceptualized as involving different kinds of changes; some require the enrichment of existing knowledge structures, and others require the creation of altogether new structures. Current discussions of the notion of restructuring in knowledge acquisition differentiate between weak and radical restructuring. Weak restructuring involves the creation of new, higher-order relations between existing concepts, whereas radical restructuring involves a fundamental change in schemata, similar to paradigm shifts in the history of science. (Vosniadou & Brewer, 1987, p. 62)

Far less is known about actual mechanisms underlying these changes than about conditions that seem to facilitate them (the bottom line for educators anyway): One encounters a situation with enough that is familiar to make it meaningful for the most part, but with unanticipated patterns or consequences. Vosniadou and Brewer (1987) suggest Socratic dialogues and analogies as pedagogical techniques to facilitate restructuring. Using them effectively requires taking into account not only the target knowledge structures, but the learner's current structures. Lesh and Lamon (1992, p. 23) describe the use of case studies in fields where the goals of instruction are associated with the construction of models for building and understanding complex systems. Relationships in the specific case are highlighted as the foundation of recurring patterns, which are then related to other specific cases to promote the construction of more general encompassing structures.

With practice, some kinds of information processing are automatized, or "compiled." Young children learn to recognize letters and words with concentration and conscious effort; practiced readers are practically unaware of individual words as they grapple with the concepts, the motivations, the implications of the texts they encounter. Similar phenomena occur in every field of learning. A skilled roentgenologist, for example, quickly identifies a spot in a chest X-ray as a tumor, although to a novice it looks like any other shadow on the plate (Lesgold et al., 1981).

Are Schemas "Real"?

The previous sections fall into the easy style of talking about schemas as if they correspond directly to something inside people's heads—not at all unlike the language of intelligence testing. But the reality of knowledge structures as we have described them is actively debated in the artificial intelligence (AI) community. Rodney Brooks's mechanical "creatures" display such complex activities as following people around a room and finding electrical outlets to recharge their batteries, using only layers of parallel primitive units that communicate in primitive ways. Although a schema theory could "explain" his creatures' behavior, Brooks (1991) emphasizes that they incorporate no central representation of concepts. In contrast, most AI researchers do explicitly code the relationships that constitute concepts into their programs, and others,

> . . . connectionists[,] seem to be looking for explicit distributed representations to spontaneously arise from their networks. We harbor no such hopes because we believe representations are not necessary and appear only in the eye or mind of the observer. (Brooks, 1991, p. 154)

For cognitive science, determining the locus and mechanisms of knowledge structures is the paramount objective. With the benefit of hindsight, our successors may see this as a guiding vision to exciting breakthroughs, or as naive a dead end as the alchemists' quest to transform lead into gold. Schema theory has today the ontological status Spearman's *g* had nearly a century ago: a conceptual tool for talking about certain patterns that seem to recur in human behavior, possibly useful for solving some practical problems. For educators, the objective is discovering when and how planning instruction in this framework helps students learn. For those of us in test theory, the objective is determining how to gather and interpret information to guide these efforts. If nothing else, the cognitive paradigm generates, where the trait paradigm could not, a common universe of discourse for learning and assessment.

Considerations for Test Theory

> Essential characteristics of proficient performance have been described in various domains and provide useful indices for assessment. We know that, at specific stages of learning, there exist different integrations of knowledge, different forms of skill, differences in access to knowledge, and differences in the efficiency of performance. These stages can define criteria for test design. We can now propose a set of candidate dimensions along which subject-matter competence can be assessed. As competence in a subject-matter grows, evidence of a knowledge base that is increasingly *coherent*, *principled*, *useful*, and *goal-oriented* is displayed, and test items can be designed to capture such evidence. [emphasis original] (Glaser, 1991, p. 26)

We must begin every application by asking "What do we want to make inferences about?" and "Why do we want to make them?" The answers should be driven by the nature of the knowledge and skills we want to help students acquire, the psychology of acquiring that knowledge, and a determination of who will use the information (teachers, parents, legislators, researchers, the students themselves) and how they will use it. There is no single "true" model for educational testing, but only models more or less useful for various purposes, by virtue of the information they convey. There is no single "best" method for gathering data, but only methods more or less effective at evoking evidence for the inferences to be made. These factors can vary dramatically across applications, with seemingly antithetical approaches sharing only a mandate to provide information consistent with a conception of how competence develops in a learning area; for example:

- *John Anderson's intelligent tutoring systems* (ITSs; see, e.g., Anderson & Reiser, 1985) characterize competence in a domain such as LISP programming as the capacity to utilize a specified set of production rules, or condition-action relationships. The tutor models a student in terms of which production rules she has mastered, and estimates her current status from the frequency with which she employs production rules in appropriate situations. These estimates are the basis of comments to the student, problem selection, and subsequent instruction. All students are expected to incorporate these production rules in their eventual model, regardless of the exact structure any student actually internalizes (e.g., understandings may be very different for student who enters already knowing FORTRAN), and only apparent production-rule usage is monitored. Thus, not all aspects of the structures of students' developing knowledge are modeled—only key aspects deemed sufficient to guide instruction and monitor targeted competencies.
- *The American Council of Teachers of Foreign Language* (ACTFL) *Proficiency Guidelines* describe stages of developing language in reading, writing, speaking, and listening (ACTFL, 1989). Table 1 contains excerpts from the reading guidelines. These generic scales are based on theories of language acquisition, as observed across languages; guidelines for specific languages help examiners map observed behavior to this more abstract frame of reference. Note the guidelines' distinction between familiar and unfamiliar contexts. Since what's familiar to one student is unfamiliar to another, the same behavior from two students can lead to different interpretations in light of additional information. Note also that the grain-size of these guidelines is too coarse for specific instructional guidance. Two Mid-Novice students, for example, might require different experiences to progress to High Novice. Finally, note that mapping behavior to the ACTFL guidelines requires judgment. We shall return in Example 2 to the problem of making abstractly stated guidelines meaningful in practice.

Whatever the paradigm, learning area, or assessment method, whenever the results affect education we are responsible for assuring that the weight and coverage of evidence are appropriate to their use. As Messick (1992, p. 2) points out, "validity, reliability, comparability, and fairness are not just measurement issues, but *social values* that have meaning and force outside of

Table 1

Excerpts from the ACTFL Proficiency Guidelines for Reading

Level	Generic Description
Novice-Low	Able occasionally to identify isolated words and/or major phrases when strongly supported by context.
Novice-Mid	:
Novice-High	:
Intermediate-Low	:
Intermediate-Mid	Able to read consistently with increased understanding simple connected texts dealing with a variety of basic and social needs. . . They impart basic information about which the reader has to make minimal suppositions and **to which the reader brings personal information and/or knowledge**. Examples may include short, straightforward descriptions of persons, places, and things, written for a wide audience. [emphasis added]
Intermediate-High	:
Advanced	Able to read somewhat longer prose of several paragraphs in length, particularly if presented with a clear underlying structure. . . . **Comprehension derives not only from situational and subject matter knowledge but from increasing control of the language.** Texts at this level include descriptions and narrations such as simple short stories, news items, bibliographical information, social notices, personal correspondence, routinized business letters, and simple technical material written for the general reader. [emphasis added]
Advanced-Plus	. . . Able to understand parts of texts which are conceptually abstract and linguistically complex, and/or **texts which treat unfamiliar topics and situations**, as well as some texts which involve aspects of target-language culture. Able to comprehend the facts to make appropriate inferences. . . . [emphasis added]
Superior	Able to read with almost complete comprehension and at normal speed expository prose on **unfamiliar subjects** and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on the knowledge of the target culture. . . . At the superior level the reader can match strategies, top-down or bottom-up, which are most appropriate to the text. . . .
Distinguished	:

Note. Based on the ACTFL Proficiency Guidelines, American Council on the Training of Foreign Languages (1989).

measurement wherever evaluative judgments and decisions are made." This is where test theory, broadly construed, comes in. It means defining what we wish to accomplish, specifying what we need to know about students to achieve it, and constructing a framework in which we can determine how well we're doing. Only then can we tell if we're succeeding, see where we are falling short, and glean clues to improve efforts to achieve our stated goals.

The paragraphs below discuss pervasive issues that arise when one attempts to frame assessment within a cognitive paradigm. The discussion is general and discursive, but each holds specific implications for models and methods in any given application. The examples that follow the discussion, therefore, will show how we grapple with some of these issues in three current projects.

The nature of the "student model." Test theory is statistical machinery for reasoning from students' behavior to conjectures about their competence, as framed in a particular conception of competence. In a particular application, this conception takes the form of a set of aspects of skill and knowledge that are important for the job at hand, be it guiding further instruction or summarizing the stages of competence students have reached. These are the variables in a "student model," as I use the term: a simplified description of selected aspects of the infinite varieties of skills and knowledge that characterize real students. Depending on the purpose, one might distinguish from one to hundreds of aspects. They might be expressed in terms of numbers, categories, or some mixture; they might be conceived as persisting over long periods of time, or apt to change at the next problem-step. They might concern tendencies in behavior, conceptions of phenomena, available strategies, or levels of development. These variables are not directly observable. We observe only students' behavior in limited circumstances—indirect evidence about competence as more abstractly conceived in the student model.

My use of the term "student model" is much broader than its typical use in AI, where "student model" usually means "runnable model," or a set of rules and conditions for their use that can be applied to provide an answer to any problem in a domain of interest (Clancey, 1986). As in Anderson's LISP tutor, this can include incomplete or erroneous rules, to mimic the behavior of students with incomplete or erroneous understandings. In my usage, this is indeed an instance of a student model; but so are "domain true scores" from CTT, status on ACTFL scales of demonstrated capabilities in second language, and categorizations of mastery of generally-stated skills in high school algebra—none "true," yet all potentially useful for certain real-world educational problems.

Obviously any student model oversimplifies the reality of cognition (whatever that may be!). In real-world educational assessment, utility is the bottom line. Greeno (1976, p. 133) points out that "it may not be critical to distinguish between models differing in processing details if the details lack important implications for quality of student performance in instructional situations, or the ability of students to progress to further stages of knowledge and understanding."[5] For immediate feedback for short-term instructional decisions, as in intelligent tutoring systems, there is a need for more detail in the student model. Enough information may be otherwise available about the student, however, that great detail is not required over a broad range of aspects of competence, but only those involved in the immediate decision. For accountability purposes, a coarser grain-size will suffice, although ideally the student model should be construed as a collapsing of a model that makes sense at the fine grain-size (see Example 1 below). Coherence of competence models in this manner allows feedback to be *consistent* with the learning model, even if it does not provide sufficient detail for all purposes under that conception.

The student's point of view. When assessment inferences are grounded in the cognitive paradigm, one must determine the extent to which the student model should reflect the student's perception of the tasks in the domain. The standard mental measurement paradigm attends to the problem stimulus only from the tester's point of view, administering the same tasks to all examinees and recording outcomes in terms of behavior categories applied in the same way for all examinees. Behavior constitutes *direct* evidence about behavioral tendencies. But in problem solving, "the search process is driven by

[5] An analog is the Smith & Wesson "Identikit," which helps police construct likenesses of suspects. Faces differ in infinitely many ways, and skilled police artists can sketch infinitely many possibilities to match witnesses' recollections. Communities that can't support a police artist use an Identikit, a collection of face shapes, noses, ears, hair styles, and so on, that can be combined to approximate witnesses' recollections from a large, though finite, range of possibilities. The payoff lies not in how close the Identikit composite matches the suspect, but whether it aids the search enough to justify its use.

[the] products of the understanding process, rather than the problem stimulus itself" (VanLehn, 1988, p. 6). Because different knowledge structures can lead to the same behavior, observed behavior constitutes *indirect* evidence about cognitive structure. Increasingly many right/wrong item responses drive the uncertainty about a student's true score to zero without providing insight into the skills and strategies she employs.

Again the guiding principle is purpose. For example, effective tutoring demands an understanding of individuals' current knowledge. Instruction based on analogy fails when students are not familiar with the context and relationships the analogy is meant to extend. The relevant questions for tutoring are not "How many items did this student answer correctly?" or "What proportion of the population would have scores lower than his?" but, in Thompson's (1982) words, "What can this person be thinking so that his actions make sense from his perspective?" and "What organization does the student have in mind so that his actions seem, to him, to form a coherent pattern?" On the other hand, behavioral summaries may suffice for monitoring progress, as long as appropriate mechanisms are in place to guide progress along the way. Coaches find it useful to chart pole vaulters' highest jumps to *track* performance, even though details of form, approach, and conditioning must be addressed to *improve* performance. Examples 1 and 2 below show how CTT and IRT, with their purpose and usage properly (re)conceived, can serve this monitoring function in ways compatible with a conception of how proficiency develops.

Compared with inference about behavioral tendencies, a chain of inference that ends with conjectures about knowledge structures has additional links, additional sources of uncertainty, that require us to work both harder and smarter. Working harder means, first, knowing how competence in the domain develops. The inferential challenges we routinely face under the standard mental measurement paradigm, such as limited information and multiple sources of uncertainty, do not disappear when interest shifts tc inference about cognitive structure. But principled reasoning now demands, *in addition* to theory about inference under uncertainty, theory about the nature and acquisition of competence in the domain. What are the important concepts and relationships students are to learn, and how do they learn them? What evidence must we see to gauge their progress, and help determine what they should do next? Working harder also means having to gather more evidence—typically not just more of the same kind of data, as in CTT and IRT, but of multiple kinds of evidence—if we want to disambiguate competing explanations of behavior (Martin & VanLehn, in press). And gathering and interpreting *direct* evidence for development over time or for productive performance requires more resources than the *indirect* evidence provided by familiar achievement tests.

It follows that working smarter first means being clear about exactly what inferences we want to make. This done, working smarter next means using strategies and techniques analogous to those long used to make inference under the mental measurement paradigm more efficacious: knowing exactly what it is we want to make inferences about, so we don't waste resources collecting *data* that hold little value as *evidence* for our needs. Identifying, then reducing, sources of uncertainty all along the chain of inference, as when training judges to use a rating scheme, or tuning tasks to evoke evi .ence about the skills of interest while eliminating extraneous sources of difficulty. Using data-capture technologies to reduce costs (e.g., Bennett, 1993, on AI scoring). Capitalizing on statistical design and analysis concepts to increase efficiencies (e.g., Shoemaker, 1975, on matrix sampling for assessing groups rather than individuals). Finally, working smarter means recognizing the role of conditionality in inference.

The role of conditionality in inference. The "traits" that achievement tests purportedly measure, such as "mathematical ability," "reading level," or "physics achievement," do not exist *per se*. While test scores do tell us something about what students know and can do, any assessment task stimulates a unique constellation of knowledge, skill, strategies, and motivation within each examinee. To some extent in any assessment comprising multiple tasks, what is relatively hard for some students is relatively easy for others, depending on the degree to which the tasks relate to the knowledge structures that students have, each in their own way, constructed. From the behavioral perspective, this is "noise," or measurement error, leading to low reliability or low generalizability under CTT, low item discrimination parameters or low person-separation indices under IRT. It obscures what one is interested in under that perspective, namely, locating people along a single dimension as to a *general* behavioral tendency; tasks that don't line up people in the same way are less informative than ones that do. (Hence the so-called "low generalizability" phenomenon often associated with performance assessment; Shavelson et al., 1992.)

These interactions are fully expected from the cognitive perspective, since knowledge typically develops first in context, then is extended and decontextualized so it can be applied more broadly to other contexts. How to handle interactions depends again on the way competence develops in the area of interest and on purpose of assessment. The same task can reveal either vital evidence or none at all, depending on the relationship of the information it carries to what is known from other sources of information. The test theory of the standard mental measurement paradigm does not addresses this principle at the level of the tasks, but at the level of the combined test score. The greater investment that each task demands and the more contextual knowledge it demands, the less efficient this approach becomes. The in-depth project that provides solid assessment information and a meaningful learning experience for the students whose prior knowledge structures it dovetails, becomes an unconscionable waste of time for students for whom it has no connection.

Consider, for example, a course that helps middle school students developing their understandings of proportionality. Each student might begin in a context with which she was personally familiar, perhaps dividing pizzas among children or planning numbers of fish for different sized aquariums. Early assessment would address each student's understanding of proportionality, *conditional on the context in which she was working*. Having everyone answer a question about the same context or about a randomly-selected context would not be an effective way to gather evidence about learning *at this stage*. Over the next few weeks, each student might carry out several investigations, eventually moving to unfamiliar contexts. Now a random sample of tasks *would* be a useful check on the degree to which each student, starting from his or her own initial configuration of knowledge, had developed a schema general enough to apply to all the contexts in the lesson. A final project might challenge students to push proportionality concepts in contexts they chose themselves. Judges would map performance in possibly quite different contexts to a common framework of meaning, rating the degree to which various aspects of understanding had been evidenced. As in the early assessment, inference at this higher level of competence would be again conditional on the context in which it has been evinced.

We can now see how such an apparently straightforward term as "difficulty" can take contradictory meanings under different paradigms. We agree that when students respond to different assessment tasks, our inferences should somehow "take the difficulty of the tasks into account." But difficulty from whose point of view? The examiner's, in ignorance about the student, or the student's? The true concern is a deeper question: Given observations in different settings, how do we draw inferences about competence as defined in such-and-such a way? The same data can have different meanings under two different conceptions of competence. The word "difficulty" may be used in each case, but to describe qualitatively different patterns among people, skills, and performances—neither right or wrong, both well-defined and useful within their respective universes of discourse:

- For inferences about overall proficiency in a domain of prespecified tasks, "difficulty" is defined empirically from the tester's point of view. Under CTT, "difficulty" means the proportions of people who would answer an item correctly. Under IRT, it means items' relative likelihoods of correct response at different levels of overall proficiency.[6] Suppose we must predict whether Mary or Charlie would correctly answer more tasks from a domain, knowing only that Mary succeeded on a task most people missed and Charlie succeeded on a task most people got right. The smart money is on Mary.

- From the cognitive perspective, "knowing a task's difficulty" means, for a given student, knowing how to interpret the evidence her performance conveys about her competence, in light of how competence develops in the domain. This may require interpretation in light with other information. For example, knowing the reader's familiarity with the content of a text is important for interpreting her behavior in terms of the ACTFL levels. Securing additional information can be explicit, such as knowing she studied Spanish, or implicit, such as knowing that this was the problem context she chose (e.g., give a talk on a topic you have done research on). In light of additional information, the same observed behavior can have different implications for different students, while different behaviors can map to the same conclusion.

For example, consider the two tasks, (a) What is the word for "pencil" in Spanish? and (b) What is the word for "pencil" in Russian? In terms of

[6] The same ideas generalize to measured responses and to ordered levels of response quality.

specified-behavior-in-a-specified-domain, the Russian task is "more difficult" for American college students simply because fewer know Russian than know Spanish. A correct answer to the Russian task "is awarded more credit" than a correct answer to its Spanish counterpart, in terms of expectations for responses to items in the domain not yet observed. But suppose that in learning both languages, "pencil" is frequently used (one of the 1000 most common words) and is introduced at a similar point in classes. If the competence of interest is "development along the ACTFL scale in *a* foreign language," then the evidence for ACTFL proficiency from either item *for an examinee studying that language* will be similar. The items are "equally difficult" from this perspective.

Examples

Example 1: Integrating Cognitive and Psychometric Models to Measure Document Literacy

> Summary test scores, and factors based on them, have often been thought of as "signs" indicating the presence of underlyıng, latent traits. . . . An alternative interpretation of test scores as samples of cognitive processes and contents, and of correlations as indicating the similarity or overlap of this sampling, is equally justifiable and could be theoretically more useful. The evidence from cognitive psychology suggests that test performances are comprised of complex assemblies of component information-processing actions that are adapted to task requirements during performance. The implication is that sign-trait interpretations of test scores and their intercorrelations are superficial summaries at best. At worst, they have misled scientists, and the public, into thinking of fundamental, fixed entities, measured in amounts. Whatever their practical value as summaries, for selection, classification, certification, or program evaluation, the cognitive psychological view is that such interpretations no longer suffice as scientific explanations of aptitude and achievement constructs. (Snow & Lohman, 1989, p. 317)

Snow and Lohman note that sometimes it really is useful to know how proficient students are in certain domains of problems, as indicated by their performance on a sample of them. But while the trait and behavioral paradigms end with statements about tendencies in behavior, a cognitive perspective can offer benefits even when we do use standard test theory to gather evidence in such applications:

- Defining and structuring the domain of tasks.
- Enriching the interpretation of scores.
- Reducing costs and gaining efficiencies.
- Improving the quality of the tasks.
- Identifying students for whom the single-number score is misleading.

This section illustrates some of this potential in a line of research being pursued by Kathy Sheehan and me. We focus here on the measure of document literacy introduced in the Survey of Young Adult Literacy (SYAL; Kirsch & Jungeblut, 1986). SYAL included 63 tasks designed to evoke the skills needed to locate and use information contained in non-prose formats such as forms, tables, charts, signs, labels, indexes, schematics and catalogs. Most of the tasks require open-ended responses. For example, respondents were directed to fill in a deposit slip, determine eligibility from a table of employee benefits, and follow a set of directions to travel from one location to another using a map. Interviewers administered the tasks to a nationally representative sample of approximately 3,600 young adults. In addition to information about responses to individual tasks, the survey was charged with providing summaries of performance in the population. To this end, an IRT model was fit, and distributions of overall proficiency in terms of an IRT variable were produced.

An item response theory (IRT) model gives the probability that an examinee will make a particular response to a particular test item as a function of unobservable parameters for that examinee and that item. Our example uses the Rasch for dichotomous items:

$$P\left(X_j = x_j | \theta, \beta_j\right) = \frac{\exp\left[x_j\left(\theta - \beta_j\right)\right]}{\left[1 + \exp\left(\theta - \beta_j\right)\right]}, \qquad (1)$$

where X_j is the response to Item j (1 for right, 0 for wrong); θ is the examinee proficiency parameter; and β_j is the difficulty parameter for Item j. Rewriting this expression as the logarithm of the odds that the respondent would respond correctly (denoted $P_{j1}(\theta)$) as opposed to incorrectly ($P_{j0}(\theta)$) focuses attention on the presumed lack of interaction between the difficulty of an item and individual respondents:

$$\ell n\left[P_{j1}(\theta)/P_{j0}(\theta)\right]=\theta-\beta_j. \qquad (2)$$

The IRT model, in and of itself, simply does not address the question of why some items might be more or less difficult than others. Fitting an IRT model is an empirical exercise, capturing and quantifying the patterns that some people tend to answer more items correctly than others, and some items tend to be answered correctly less often than others. The conception of document literacy competence embodied by the IRT model is simply the tendency to perform well in the domain of tasks.

From a cognitive perspective, what makes a task difficult for a particular individual is the match-up between her knowledge structure and the demands of the task. As discussed above, these match-ups can vary substantially from one person to another for any given task. An IRT item difficulty parameter captures only a population-level characteristic, the relative ordering of items *on the average*. The summaries of the difficulties of items and the proficiencies of persons that the IRT parameters embody miss information to the extent that items are hard for some people and easy for others.

It is sometimes possible, nevertheless, to characterize tasks from an expert's point of view—that is, in terms of the knowledge, operations, and strategy requirements, and working memory load of an ideal solution. One may thus gain insights into the features of tasks that tend to make them relatively easy or hard in a population of examinees. For example, Scheuneman, Gerritz, and Embretson (1991) accounted for about 65% of the variance in item difficulties in the Reading section of the National Teacher Examination (NTE) with variables built around syntactic complexity, semantic content, cognitive demand, and knowledge demand.

Scheiblechner (1972) and Fischer (1983) integrated such cognitive information into IRT with the Linear Logistic Test Model (LLTM), which models Rasch item difficulty parameters as linear functions of effects that correspond to key features of items. Mislevy (1988) extended the LLTM to allow for variation of difficulties among items with the same key features, by incorporating a residual term to yield

$$\beta_j = \sum_{k=1}^{K} q_{kj} \eta_k + \varepsilon_j, \tag{3}$$

where η_k is the contribution of Feature k to the "difficulty" of an item, for k=1,...,K item features; q_{kj} is the extent to which Feature k is represented in Item j; and ε_j is a N(0,ϕ^2) residual term, with (estimated) variance ϕ^2.

Sheehan and Mislevy (1990) implemented this model with item features from Mosenthal and Kirsch's (1991) cognitive analysis of the difficulty of document literacy tasks. Their system begins by characterizing the information contained in documents and document task directives according to three basic levels of organization: (a) the organizing category, (b) the specific category, and (c) the semantic feature. Semantic features are bits of information that belong to specific categories, which are nested within distinct organizing categories. Specific categories can also be nested within other specific categories; complex documents can have several levels of nested specific categories. Using this characterization, Kirsch and Mosenthal defined three classes of variables they expected to correlate with task difficulty: (a) variables that characterize the length and organizational complexity of the *materials* which document tasks refer to; (b) variables that characterize the length and organizational complexity of task *directives*; and (c) variables that characterize the difficulty of the task solution *process*. They are listed in Table 2.

These features accounted for about 80% of the variance of the IRT task difficulty parameters (β). The structural complexity of material and directives were important factors, but the highest contributions were associated with process variables. The details of such analyses help item writers control the difficulty of the tasks they develop. No items in this study were exceptionally easier or harder than their modeled features would suggest. Such outliers would direct item writers' attention to tasks that might be unexpectedly difficult for irrelevant reasons, or unexpectedly easy because of unintended cues.

Table 2

Task Features Codings for the Document Literacy Model

Materials variables

(1) The number of organizing categories in the document;

(2) The number of organizing categories in the document that are embedded;

(3) The deepest level of embedding for an organizing category;

(4) The number of specific categories in the document;

(5) The number of specific categories in the document that are embedded; and

(6) The deepest level of embedding for a specific category.

Directive variables

(1) The number of organizing categories in the directive;

(2) The number of organizing categories in the directive that are embedded;

(3) The deepest level of embedding for an organizing category;

(4) The number of specific categories in the directive;

(5) The number of specific categories in the directive that are embedded; and

(6) The deepest level of embedding for a specific category.

Process variables

(1) Degree of Correspondence. This variable is scored on a one-to-five integer scale. It indicates how explicitly the information requested in the directive or question matches the corresponding information in the text, with higher values indicating less explicit correspondence and therefore, more difficulty. For example, tasks requiring a single literal match are scored one, tasks requiring an inferential text-based match are scored three, and tasks requiring matches based on specialized prior knowledge are scored five.

(2) Type of Information. This variable concerns the type and number of restrictive conditions that must be held in mind when identifying and matching features. Lower values on a one-to-five scale signify less restrictive conditions.

(3) Plausibility of Distractors. Document tasks typically require the examinee to skim an entire document to locate a piece of requested information. Since any piece of information in the document could be interpreted as the requested information, document task "distractors" include all pieces of information embedded in the document. The degree of plausibility of a distractor is measured by the extent to which the information embedded in the document shares semantic information with the correct answer to the question or directive. This variable is scored on a one-to-five scale, with lower numbers indicating less shared semantic information and higher numbers indicating more.

Note. Based on the analysis of Mosenthal and Kirsch (1991).

The location of items along the Rasch IRT proficiency scale is directly related to the measures of individuals' proficiencies: items' β values indicate the probabilities of success from people at given levels of θ. Modeling the locations of tasks with particular configurations of processing requirements on this scale indicates what a person at a given level of IRT proficiency might be expected to do in terms of requirements of tasks—a probabilistic link between empirical IRT summaries of observed response and cognitive explanations. While the IRT θ still only captures overall competence, this connection adds a layer of meaning to score interpretation.

Recall, however, that this modeling is just "on the average." It only relates the cognitive model to an analytic model that posits the items line up in the same way for everyone. To some degree, what is easy for one person will be hard for another. This interaction, missing from the IRT summary, can be accessed through analyses of residuals from the model's fit. The same processing-feature structure can be used to examine unexpected response patterns of individual respondents, complementing overall-proficiency θ estimates with diagnostic information.

We are now exploring the extent to which cognitive requirements (and other sources of information about tasks) provide information about IRT item parameters in a variety of applications. Even if the IRT paradigm is sufficient for summarizing and monitoring purposes, exploiting information from the cognitive perspective can reduce or even eliminate pretesting meant to estimate item parameters (Mislevy, Sheehan, & Wingersky, 1993). This opens the door to using IRT with tasks created on the spot with generative algorithms founded upon cognitive processing models (Bejar, 1993; Irvine, Dann, & Anderson, in press).

Example 2: AP Studio Art Portfolios

> As compared to measurement, assessment is inevitably involved with questions of what is of value, rather than simple correctness. Questions of value require entry and discussion. In this light, assessment is not a matter for outside experts to design; rather, it is an episode in which students and teachers might learn, through reflection and debate, about the standards of good work and the rules of evidence. (Wolf, Bixby, Glenn, & Gardner, 1991, pp. 51-52)

Performance assessment commands attention partly because it provides direct evidence about productive aspects of knowledge, and partly because of its potential impact on educational practice—"What you test is what you get" (Resnick & Resnick, 1989). A distinguishing characteristic of performance assessment is that the student's response is no longer simply and unambiguously classified as right or wrong; judgment is required *after* the response has been made. Performance assessment raises a new set of inferential issues, some of which have counterparts in multiple-choice testing, but others for which we have much experience. Table 3 lists some of these issues.[7] Many highlight dimensions along which performance assessment systems and scoring systems vary. We need to learn more about the consequences, costs, characteristics, advantages, type of evidence provided, and so on, of these alternatives, so that we can construct performance assessment systems that provide the right *kind* of evidence for a given purpose, with the required *weight* and *coverage* of evidence, expending the right level of *resources.*

In most performance assessments, judgment is the crucial link in the chain of reasoning from performance to inference about students. As with our opening example of directed graph for checker games, each of several tasks may, in and of itself, stimulate the kind of creative or problem-solving thinking we are interested in—to no avail unless we can distill from the performance the critical evidence for the targeted inferences. It is thus essential to establish a common framework of meaning among readers, shared standards for recognizing what is important in performance and mapping it into a summarizing structure. It is no less essential that the same framework of meaning be common to students and teachers as well. Quite aside from the important issue of fairness—and students *should* know the criteria by which they will be evaluated—learning the framework for evaluation can be an essential part of learning what a course is supposed to teach, namely, the characteristics of valued work (Wolf, Bixby, Glenn, & Gardner, 1991). The instructional value of an evaluation scheme appears positively in cost/benefit equation, along with more familiar characteristics such as inter-reader agreement (Frederiksen & Collins, 1989).

[7] This listing was prepared by Drew Gitomer, Carol Myford, and myself.

Table 3

Some Inferential Issues in Performance Assessment

What is the student model? At one extreme, we can have a student model in performance assessment quite analogous to that of multiple-choice assessment: We'd set up categories, and model the student in terms of tendencies to behave in those ways across contexts. Alternatively, we could have judges interpret behaviors on a more abstract scale, such as the "levels of developing expertise in a content area." An example of the latter is the American Council of Teachers of Foreign Languages (ACTFL) generic rubrics for reading, speaking, listening, and writing, based on a functional model of language development.

High inference vs. low inference scoring. Low inference scoring systems summarize behaviors into easily agreed-upon categories, while high inference systems summarize observations in terms of more abstractly defined qualities. We see a trade-off in teacher assessment: judges agree more closely in identifying behaviors, but feel that high-inference interpretations in terms of "teaching-related traits" or "characteristics of teaching interactions" are more closely tied to conceptions of competence.

Generalizability. Performance assessment tasks typically take more time than, say, multiple-choice items. There is a tradeoff between the depth of information we can obtain in a given context, and how broadly we can look at different contexts. What types of skills and purposes call for depth? For some purposes, should we construct assessments that evoke a combination of types of evidence, in order to learn something about depth and breadth?

Norm-referenced vs. criterion-referenced scoring. An example: In ARTS PROPEL, 8th- and 12th-graders' writing portfolios are rated on the same 1-5 scales for a variety of characteristics. Should ratings of 8th-graders' work take into account that they are 8th graders, so the same portfolio would receive lower ratings if were produced by a 12th grader? Or should the meanings of the rating points be identical over grades? Both are options, and each focuses information better for different purposes. One conclusion is, though, that all of the judges should agree on how they are using the scale in this respect.

How can the meanings of scoring systems be communicated? Communication is required among judges, certainly, for reliability; it seems equally important to communicate criteria to students as well. The standard setting processes the College Board's Advanced Placement (AP) program use to train judges works back and forth between examples and verbal rubrics. Are different approaches better suited to different settings and different purposes? How about when the very process of learning to understand evaluation criteria is an essential part of developing competence in the domain? Does this weigh in favor of scoring systems with educational pluses, even if they entail less agreement among raters?

Local vs. central scoring. Most current models have a core of central judges. To handle portfolios from tens of thousands of students, California envisages local teacher ratings, perhaps within a system of cross-validation and moderation. What tradeoffs are involved with these models? What other models might there be? How do we match characteristics of models with purposes, resources, and systemic consequences?

Product vs. process. It is often easier to judge products of performance assessment tasks that yield products, than to judge the processes by which they were produced. For what purposes, with what kinds of student models, should we judge process, and how do we best do so?

When do we need to know context and/or intentions? When is it necessary, to evaluate the evidence a performance conveys, to take context into account? Context can include student background as to education or culture. Context can be internal to the student as well: How does

Table 3 (continued)

the student see the task, and how is he or she trying to accomplish it? The ACTFL rubrics distinguish between texts familiar and unfamiliar to the examinee, and we need to know about the relationship between the task and the individual student in order to draw inferences from behavior to levels on the ACTFL scales.

What are the implications of *choice*? Assessing skills that develop and are evidenced only in context becomes tricky when the context varies from student to student. Often students themselves possess insight into the contexts that allow them to demonstrate what they can do. Doctoral dissertations are an example. We miss the point if we require all students to write dissertations on the same topic, or assign topics at random. We need to develop inferential models to deal with evidence from performances in which students have varying kinds and degrees of choice about what they will do.

The observer as a filter. Judges, as unique individuals, will see and interpret performances in ways influenced by their own history, experiences, and values. How do we maximize the extent to which they agree on what to look for and how to interpret it? How do we help them understand their perspectives, and the impact on the judging processes? By what statistical methods can we detect unusual judge/performance interactions?

Characteristics of different type of rubrics. Rubrics can be described as holistic, analytic, or interpretive. They can be generic—meant to be used with any of a family of tasks—or task specific. What kinds of rubrics are suited to different purposes and situations? What are tradeoffs as far as consistency, ease of learning, etc.? One might choose a generic rubric for an assessment meant to span over time and across many tasks, for example, in preference to more-easily-agreed-upon specific rubrics that provide evidence that is hard to connect.

Statistical machinery for analysis, summarization, and quality control. Too many performance/judgment interactions will take place in most systems for a single person to observe and evaluate. A mechanism is needed to bring together summary information for the purposes of summarizing key aspects of the operation of the system as a whole, highlighting aspects which might be improved by changing the system or the judge training, and flagging anomalies in specific performance/judgment interactions that need attention. Traditional generalizability analyses provide some of this, but are not adequate for all needs and purposes. Latent variable models that model individual task and judge effects (e.g., Linacre's (1989) FACETS analysis) are a promising route.

Linking results from different tasks. Ties in with the aforementioned issues of generalizability, statistical machinery, the nature of rubrics, and judge training. If different students respond to different tasks at different points in time, how do we interpret evidence in a common frame of reference? E.g., a common generally-phrased rubric, with alignment largely through the mechanism of shared meanings, as opposed to task-specific rubrics, linked through statistical mechanisms and overlapping data.

Implications of assessment choices for the system. Practically all of the choices discussed above have implications for the educational system in which the performance assessment is taking place. They may have more or less impact, for better or worse, on students, teachers, administrators, parents, and society. What are they, and how do we evaluate them? To what degrees and in which contexts do they weigh into cost/benefit analyses of developing a performance assessment system?

Note. Based on an internal memorandum by Drew Gitomer, Carol Myford, and Robert Mislevy.

Carol Myford and I are working with the College Board's Advanced Placement (AP) Studio Art program to explore issues in monitoring and improving inference in performance assessment. Advanced Placement assessments are meant to determine whether high school students exhibit knowledge and skills commensurate with first-year college courses in a content area. AP Studio Art is one of the nation's longest extant portfolio rating systems. Students develop works during the course of the year, through which they demonstrate the knowledge and skills described in the AP Studio Art materials. The portfolios are rated centrally by artist/educators at the end of the year, using standards set in general terms and monitored by the AP Art advisory committee. At a "standards setting session," the chief faculty consultant and table leaders select portfolios to exemplify the committee's standards. The full team of about 25 readers spends the equivalent of one day of the week-long scoring session examining, discussing, and practicing with these and other examples to establish a common framework of meaning. Aspects of the assessment include ratings on three distinct sections of each portfolio, multiple ratings of all sections for all students, and virtually unbridled student choice in demonstrating their capabilities and creative problem-solving skills, within the guidelines set forth for the sections.

Students may elect to participate in two types of portfolio assessment in AP Studio Art, Drawing and General Art. We address General Art. Among the requirements for each portfolio are four works submitted in their original form; eight slides that focus on color and design, eight slides of drawings, and four of three-dimensional work; and up to 20 slides, a film, or a videotape illustrating a concentration on a student-selected theme. These requirements ensure that evidence about key aspects of artistic development will be evoked—although the wide latitude of choice of medium and expression virtually guarantees that the particular form the evidence takes will vary considerably from one student to another. We have focused on Section A, the four works submitted in original form to be rated as to "overall quality," and Section B, the student's "concentration," the up-to-20 works mentioned above and a paragraph or two describing the student's goals, intentions, influences, and other factors that help explain the series of works.

The AP Studio Art portfolio assessment reveals the contrast between "standardized" and "nonstandardized" assessments as a false dichotomy, a hindrance as we develop broader ranges of assessment methodologies. Any assessment might be implemented in countless ways; there could be differences, small or large, as to tasks, administration conditions, degree of student choice, availability of resources, typeface, identity and number of judges, and so on. *Standardizing* an aspect of an assessment means limiting the variation that students encounter in that aspect as a way of sharpening the evidence about *certain* inferences from what is observed, while perhaps simultaneously *reducing* evidence about others. Did Duanli score higher than Marilyn because she had more time, easier questions, or a lenient grader? Standardizing timing, task specifications, and rating criteria reduce the chance that this was so; it simultaneously reduces information about the differential settings in which they might do best. As in AP Studio Art, assessing students' developing competence when there is neither a single path toward "better" nor a fixed and final definition of "best" may require different kinds of evidence from different students (Lesh, Lamon, Lester, & Behr, 1992, p. 407). Questions about which aspects of an assessment to standardize to what degrees arise under all purposes and modes of testing, and under all views of competence. Answers depend on the evidential value of the observations in view of the purposes of the assessment, the conception of competence, and the requisite resource demands.

Our study uses two distinct perspectives, "statistical" and "naturalistic," which we believe are required in tandem to analyze and improve a system the size of AP Studio Art—currently some 7000 portfolios x 5 rating areas in each portfolio x 2 or 3 ratings for each, totaling over 50,000 judgments! The statistical component reflects recent thinking about quality control in industry (e.g., Deming, 1980). One begins by establishing a statistical framework for analyzing data, to quantify typical and expected sources of variation (in our case, students, readers, and sections of the portfolios). Variability is present in any system; within a statistical framework, typical ranges can be modeled. For a system that is "under statistical control," sources of variability are identified and observations tend to follow regular patterns. Modeling these patterns is useful first because it quantifies the uncertainty for final inferences (in our case, students' final ratings on a 1-5 scale) associated with steps or aspects of the process, which can be monitored when the system is modified. Secondly, the framework highlights observations that lie outside the usual ranges of variability, often due to special circumstances that can be accommodated within the existing system or which may suggest changes to the system. It is simply impossible for any one individual to become intimately familiar with all 50,000 separate rating processes. This framework helps focus attention where it is most needed.

For the statistical component of our project, we are using Linacre's (1989) FACETS model, a generalization of Masters' (1982) PARTIAL CREDIT item response theory model. FACETS provides a statistical model for ordered-category scores, as functions of parameters for examinees, readers, tasks, and other "facets" of the observation setting that may be relevant, such as reader background and time of day. This model extends the regularity patterns embodied in IRT beyond the "tendency for specified behavior on specified tasks" paradigm in the following sense: Whereas IRT was invented to model regularities in examinees' overt behavior in contexts considered invariant over people, FACETS uses similar mathematical structures to model regularities in readers' application of common standards to possibly quite different behaviors in different contexts. In 1992, one student's concentration focused on "angularity in ceramics," while another's dealt with an "application of techniques from traditional oriental landscapes to contemporary themes." It would be easier to compare students' performances if everyone were required to work with angularity in ceramics, but that would provide no evidence about a crucial aspect of development as an artist, namely conceptualizing and confronting one's own challenges.[8]

Mathematically, the FACETS model is an extension of the simple Rasch model shown in Equation 2. The logarithm of the odds that a portfolio section with a "true" measure of θ will receive from Judge j a rating in Category k as opposed to Category k+1 on a scale with K ordered categories is given as

[8] If we were to randomly assign multiple concentrations to each student, we could learn about the interrelationships among them (in principle—remember that it takes a whole year to do just one concentration!). They might well be quite modest, indicating a "low generalizability" problem from the mental measurement perspective in which the target inference would be how one would perform on the domain of potential concentrations as a whole. But the point is that how well the ceramics student would have done with oriental landscapes is irrelevant to the inference we are really interested in. What really matters, and what we must check the quality of, is our inference about the more abstractly defined qualities that should be evinced in any student's chosen concentration.

$$\ell n\left[P_{j,k}(\theta)/P_{j,k+1}(\theta)\right] = \theta - \xi_j + \tau_k, \tag{4}$$

where ξ_k is the "harshness" parameter associated with Judge j and τ_s , for s=1,...,K, is a parameter indicating the relative probability of a rating in Category s as opposed to Category s-1. An analysis of ratings of concentrations only would have no repeated observations of students at all, but would focus on patterns among the ratings of different readers. One could, by extending the model further, explore whether students' performances across the sections of their portfolios did function as repeated observations of a single variable; that is, whether students tended to score well or poorly across the board. Then the FACETS model would include an additional term for portfolio section, say ηh:

$$\ell n\left[P_{h,j,k}(\theta)/P_{h,j,k+1}(\theta)\right] = \theta - \xi_j + \tau_k + \eta_h. \tag{5}$$

In essence, FACETS fits a main-effects model to log-odds of ratings. Variation among portfolios, as a main effect, is anticipated. These are estimates of portfolio "measures," or estimates of values disentangled from the effects of specific readers. Variation among readers, as a main effect, is not desirable. It indicates that some readers tend to be more harsh or lenient than others, no matter which portfolio they are rating. The uncertainty this entails for final ratings can be reduced by improving feedback on the application of standards to individual readers or in training sessions, or by adjusting scores for individual readers. We found little variation of this type in the 1992 Studio Art data, alleviating concerns about systematic differences between readers from secondary and higher-education settings, with more or less experience as an art educator, or with more or less experience as an AP reader. Variation at the level of readers-by-portfolios, as indicated by residuals from the main-effects model, is also undesirable but *cannot* be adjusted away by statistical means when a reader rates a section only once. It may be reduced by such means as improving reader training, sharpening the definition of standards, or distinguishing aspects that should be rated separately. Presaging the "naturalistic" component of our project, FACETS highlights particular reader/portfolio combinations that are especially unusual in view of the main effects.

Statistical analyses can tell us where to focus attention, but they can't tell us what to look for. These cases are unusual precisely because the expected causes of variation do not explain them. For example, a harsh reader's rating of 1 on a portfolio that receives 1's and 2's from other readers is not surprising; a lenient reader's rating of 1 for a portfolio that receives mostly 3's and 4's is. Further insight requires information outside the statistical framework, to seek new hypotheses for previously unrecognized factors. Such investigations constitute the "naturalistic" aspect of our project. We identified 9 portfolios each for Section A and Section B that received highly discrepant ratings from two readers. (Currently, all such occurrences are identified and rectified by a final rating from the chief faculty consultant.) We discussed each of these portfolios with two experienced readers to gain insights into the judging process in general, and into the features that made rating these particular portfolios difficult. Table 4 samples excerpts from these discussions. Several avenues for possible exploration were suggested, including the following: continued development of verbal rubrics, particularly as a learning tool for new readers; having students write statements for color and design sections, as for concentrations, to help readers understand the challenges the students were attacking; and refining directives and providing additional examples for Section B to clarify to both students and readers the interplay between the written and productive aspects of a concentration.

The attractive features of performance assessment include the potential for instructional value and the elicitation of direct evidence about constructive aspects of knowledge. Outstanding concerns include the weight of evidence it provides and the question of accountability. The approach described above addresses aspects of both concerns, for only by working back and forth between statistical and naturalistic analyses can a common framework of meaning be established, monitored, and refined over time. This study illustrates one approach, using ideas originally developed under the mental measurement paradigm but extended to a cognitive/developmental paradigm, to characterize the weight of evidence about target inferences and to provide information to increase the weight of evidence. By making the materials and results of such a process public, one can assure parents and legislators of the meaning and value of the work such assessments evoke, and of the quality of the processes by which evidence about students' competence is inferred.

Table 4

Excerpts From Discussions With AP Studio Art Judges

A hypothesis for discrepant ratings for a concentration about angularity in ceramics

R: What we've done is selected portfolios that are problematic, and this is one of them. What do you guess is problematic about this?

J: I think it's problematic . . . looking at something three dimensional in a two-dimensional format and not being able to see any of the original work coming in that is three dimensional. Especially, with some of them, you're not seeing more than one angle. . . . Now, we did have a problem with one today in the gymnasium where there were photographs of three dimensional of four pots—they weren't any of these. They were beautiful pots—the photographs weren't—but that could have been a 1 and a 4.

On the nature of the task of rating AP portfolios

R: See, the portfolio doesn't work unless you have the criteria set before, because otherwise you just become judges. And a judge goes on, if it's an open show, he picks the things he likes with only his criteria. But if the criteria are set for the portfolio ahead of time, we have to subjugate that—our criteria. And come in, and the criteria for this is this: By the end of a college freshman year in a college entrance program, is this the work that would come out? And it wouldn't be high level if it were to come out. There would be statements like this. This one I don't think would come out. I think the drawing in here is too weak to come out. . . .

S: . . . When there's a discrepancy, when we're reviewing like openly here at the table, and once in a while someone will say, "Well, gee, I gave that a 2," and someone else will say, "It's a 4," and we'll talk it out, and then one or the other person begins to see that their own maybe personal opinion has pervaded their judgment in such a way that they've been persuaded by something other than this kind of ultimate sense of . . .

R: It isn't like voting.

S: It isn't like voting, but then suddenly the light has been shined on that part of themselves that might be slightly tainted or biased by their preference for or against something, and that's when you see someone sort of become persuaded that they may have been biased in their opinion. Not that they've changed all they felt about it, but they can now raise themselves up to less self-involved—step back and see it more clearly having heard other people's verbalization. . . .

R: And it's usually bringing back. And it isn't making them change their mind about the way they feel, but it's usually bringing them back to focus on what this . . .

Dealing with "uneven" sections

E: It's always troublesome, I think, when there's weak pieces and then some very, very strong pieces, you know. I just really sort of sort them out and try to get an overall idea of what the student's doing. I sort of wonder why some of them are in there. It's a perennial question.

R: But then you have to weigh the other body of work, and if it warrants consideration against the weak pieces, then it can negate the effect of the weak pieces.

Table 4 (continued)

On rating works on topics or media one personally reacts to negatively

B: Something that you're, in a sense, sick of seeing, or something that you maybe have been taught is not proper subject matter or style, or whatever. With the kind of open arena that we have for art today, a true artistic statement can be made in any number of ways, so I try to let myself be open and aware of what this person might be trying to do. And if they are doing it well within that certain style, whether it's something I quote "approve of" or not. And, well for instance, one of the styles that comes up frequently in these ratings is a comic book, a cartoon style, because the males in this age bracket are really infatuated with drawing super heroes or metal mania, or something like that. So you see a lot of it. I tend to spend more time looking at those than I do at some of the other portfolios. Not most of them, but then some of the others are kind of questionable.

J: You try to bring all your aesthetics.

B: And I say, "Are they using the space well? Are they being inventive with how they're doing this, or are they merely mimicking somebody and not doing a very good job of it?"

J: You almost go down a whole checklist of descriptors—line, space, color, texture.

B: I kind of drill myself. I put myself under the spot light.

J: And how do those apply to the rhythm, the balance, the content? Is there harmony, unity here?

R: So we're back to the formal stuff again?

J: Right, but I think that's what we have to rely on are those formal elements and principles.

Dealing with "uneven" sections

E: . . . If half of them are excellent and the other half are not, how come? If they did four pieces of art work, how can two be so great and the other two they selected. We don't know. Did the teacher give them some bad advice? Did the teacher let the student just do it on his own? I mean, I've had experiences where the student just loved this piece of work, and it was not that good. But they just loved it, loved it. And when push came to shove, you know, I said my advice is this but you're the one that's paying $65.00, and you're the one that's going to put those four in there. And I want it to be, you know, you feel that whatever your grade was is the one that you really earned because that's what you wanted in there. And that's the way the final decision needs to go.

P: I think when you have four pieces and one of them is truly bad, it's easy not to see that piece. The other three carry the four. I think it's hard when you've got two and two, and sometimes even harder where you have one that just really knocks your socks off, and then you have three that you think, "I don't think this same person did this."

E: That's when you want that phone again.

P: Yeah, you want to say, "Who? What? Who? How?" . . . [But] I go with the benefit of the doubt towards the student. Say I have one that's outstanding and three that are really, really horrible. To me, that is a good, strong, high 2. Other people might disagree, but we are here for the kids. That's that teacher behind me. I go for the student. I tend to think that he got poor advice somewhere along the line.

Example 3: Mixed Number Subtraction

The form of the data in this final example is familiar—right/wrong responses to open-ended mixed-number subtraction problems—but inferences are carried out in terms of a more complex student model suggested by cognitive analyses. The model is aimed at the level of short-term instructional guidance. It concerns which of two strategies students apply to problems, and whether they can carry out the procedures that problems require under those strategies. While competence in domains like this can be modeled at a much finer grain-size (e.g., VanLehn's 1990 analysis of whole-number subtraction), the model in this example does incorporate the fact that the "difficulty" of an item depends on the strategy a student employs. Rather than discarding this interaction as noise, as CTT or IRT would, our model exploits it as a source of evidence about a student's strategy usage.

The data and the cognitive analysis upon which the student model is grounded are due to Kikumi Tatsuoka (1987, 1990). The middle-school students she studied characteristically solved mixed-number subtraction problems using one of two strategies:

Method A: Convert mixed numbers to improper fractions, subtract, then reduce if necessary.

Method B: Separate mixed numbers into whole number and fractional parts, subtract as two subproblems, borrowing one from minuend whole number if necessary, then reduce if necessary.

We analyzed 530 students' responses to 15 items. As shown in Table 5, each item was characterized in terms of which of seven subprocedures were required to solve it with Method A and which were required to solve it with Method B. The student model consists of a variable for which strategy a student uses, and which of the seven subprocedures he is able to apply. The structure connecting the unobservable parameters of the student model and the observable responses is that ideally, a student using Method X (A or B, as appropriate to that student) would correctly answer items that under that strategy require only subprocedures the student has at his disposal (see Falmagne, 1989; Haertel & Wiley, 1993; Tatsuoka, 1990). However, sometimes students miss items even under these conditions (false negatives), and

Table 5

Skill Requirements for Fractions Items

			If Method A used				If Method B used			
Item #	Text	1	2	5	6	7	2	3	4	5
4	$3\frac{1}{2} - 2\frac{3}{2} =$	x			x		x	x	x	
6	$\frac{6}{7} - \frac{4}{7} =$	x								
7	$3 - 2\frac{1}{5} =$	x		x	x		x	x	x	x
8	$\frac{3}{4} - \frac{3}{8} =$	x								
9	$3\frac{7}{8} - 2 =$	x	x	x	x	x		x		
10	$4\frac{4}{12} - 2\frac{7}{12} =$	x	x		x		x	x	x	
11	$4\frac{1}{3} - 2\frac{4}{3} =$	x	x		x		x	x	x	
12	$\frac{11}{8} - \frac{1}{8} =$	x	x				x			
14	$3\frac{4}{5} - 3\frac{2}{5} =$	x			x			x		
15	$2 - \frac{1}{3} =$	x	x	x				x	x	x
16	$4\frac{5}{7} - 1\frac{4}{7} =$	x	x		x			x		
17	$7\frac{3}{5} - \frac{4}{5} =$	x	x		x			x	x	
18	$4\frac{1}{10} - 2\frac{8}{10} =$	x	x		x	x	x	x		
19	$7 - 1\frac{4}{3} =$	x	x	x	x	x	x	x	x	x
20	$4\frac{1}{3} - 1\frac{5}{3} =$	x	x		x	x	x	x	x	

Skills:

1. Basic fraction subtraction
2. Simplify/Reduce
3. Separate whole number from fraction
4. Borrow one from whole number to fraction
5. Convert whole number to fraction
6. Convert mixed number to fraction
7. Column borrow in subtraction

sometimes they answer items correctly when they don't possess the requisite subprocedures by other, possibly faulty, strategies (false positives). The connection between observations and student-model variables is thus probabilistic rather than deterministic.

Inference in complex networks of interdependent variables is a burgeoning topic in statistical research, spurred by applications in such diverse areas as forecasting, pedigree analysis, troubleshooting, and medical diagnosis (e.g., Lauritzen & Spiegelhalter, 1988; Pearl, 1988). Inference networks exploit conditional independence relationships. Current interest centers on obtaining the distributions of selected variables conditional on observed values of other variables, such as likely characteristics of children of selected animals given characteristics of their ancestors, or probabilities of disease states given symptoms and test results. If the topology of the interconnections is favorable, such calculations can be carried out in real time in large systems by means of strictly local operations on small subsets of interrelated variables ("cliques") and their intersections. Lauritzen and Spiegelhalter (1988), Pearl (1988), and Shafer and Shenoy (1988) discuss updating strategies, a kind of generalization of Bayes theorem.[9] Béland and Mislevy (1992), Martin and VanLehn (in press), Mislevy (in press), and Mislevy, Yamamoto, and Anacker (1992) show how inference networks can be applied to problems in cognitive diagnosis.

Figure 2 depicts the structural relationships in an inference network for Method B only. Nodes represent variables, and arrows represent dependence relationships. The joint probability distribution of all variables can be represented as the product of conditional probabilities, with a factor for each variable's conditional probability density given its "parents." Five nodes represent basic subprocedures that a student who uses Method B needs to solve various kinds of items. Conjunctive nodes, such as "Skills 1 & 2," represent, for example, either having or not having *both* Skill 1 and Skill 2. Each subtraction item is the "child" of a node representing the minimal conjunction of skills

[9] Calculations for the present example were carried out with Andersen, Jensen, Olesen, and Jensen's (1989) HUGIN program and Noetic System's (1991) ERGO.

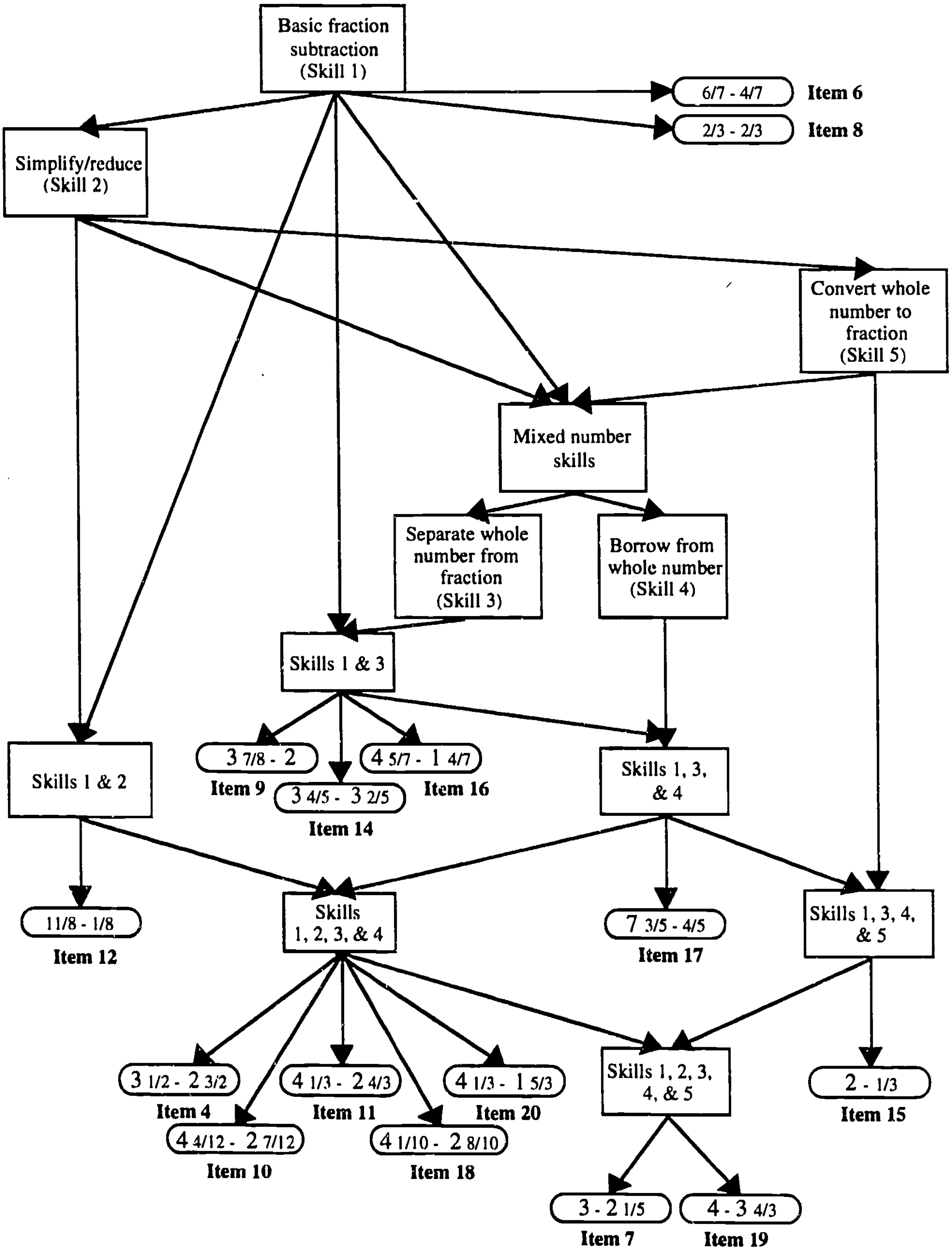
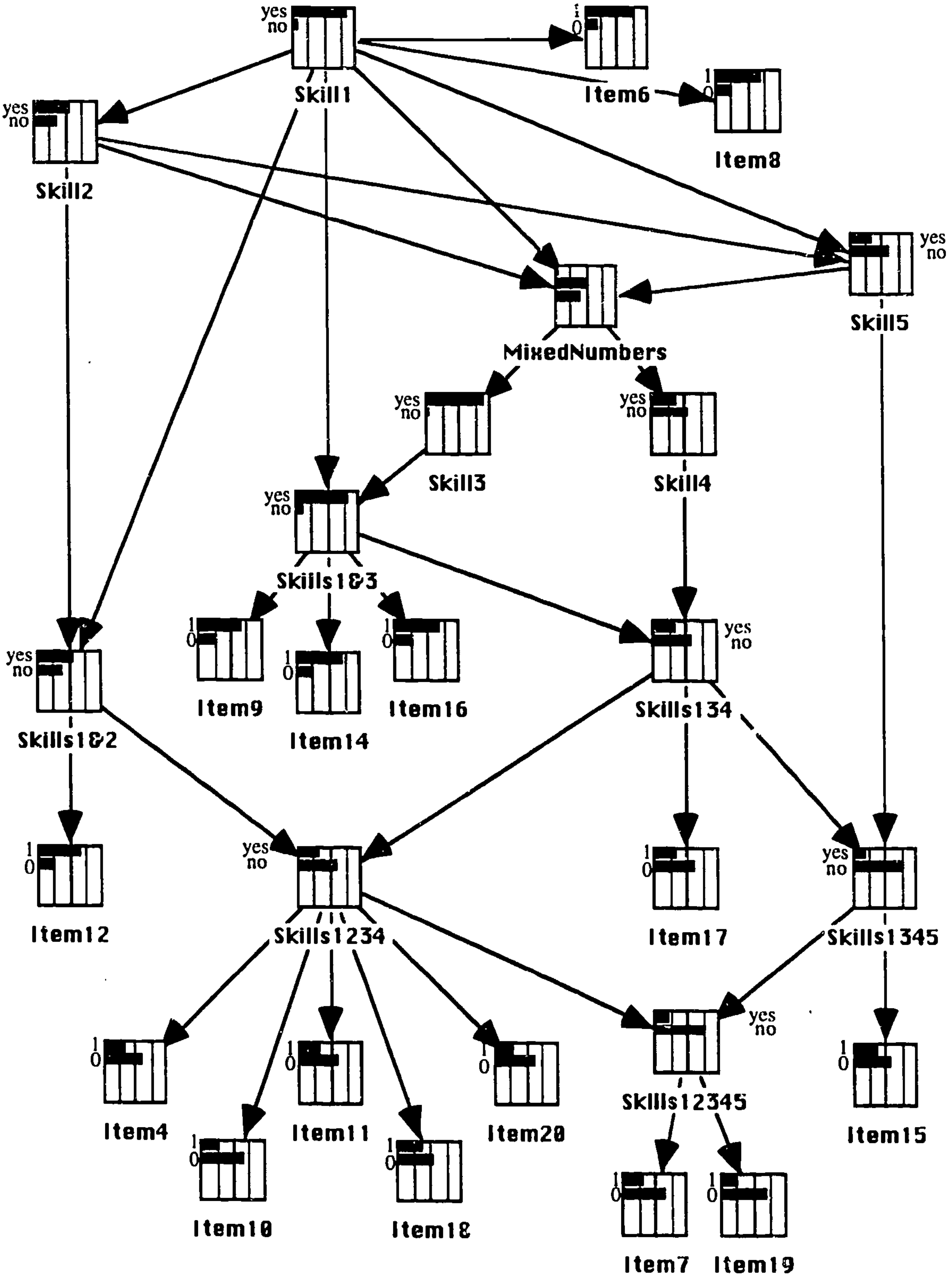


Figure 2. Structure of inference network for Method B.

needed to solve it with Method B. The relationship between such a node and an item incorporates false positive and false negative probabilities. Cognitive theory inspired the *structure* of this network; the *numerical values* of conditional probability relationships were approximated with results from Tatsuoka's (1983) "rule space" analysis of the data, with only students classified as Method B users. (Duanli Yan and I are working on estimating conditional probabilities in this network with the EM algorithm.)

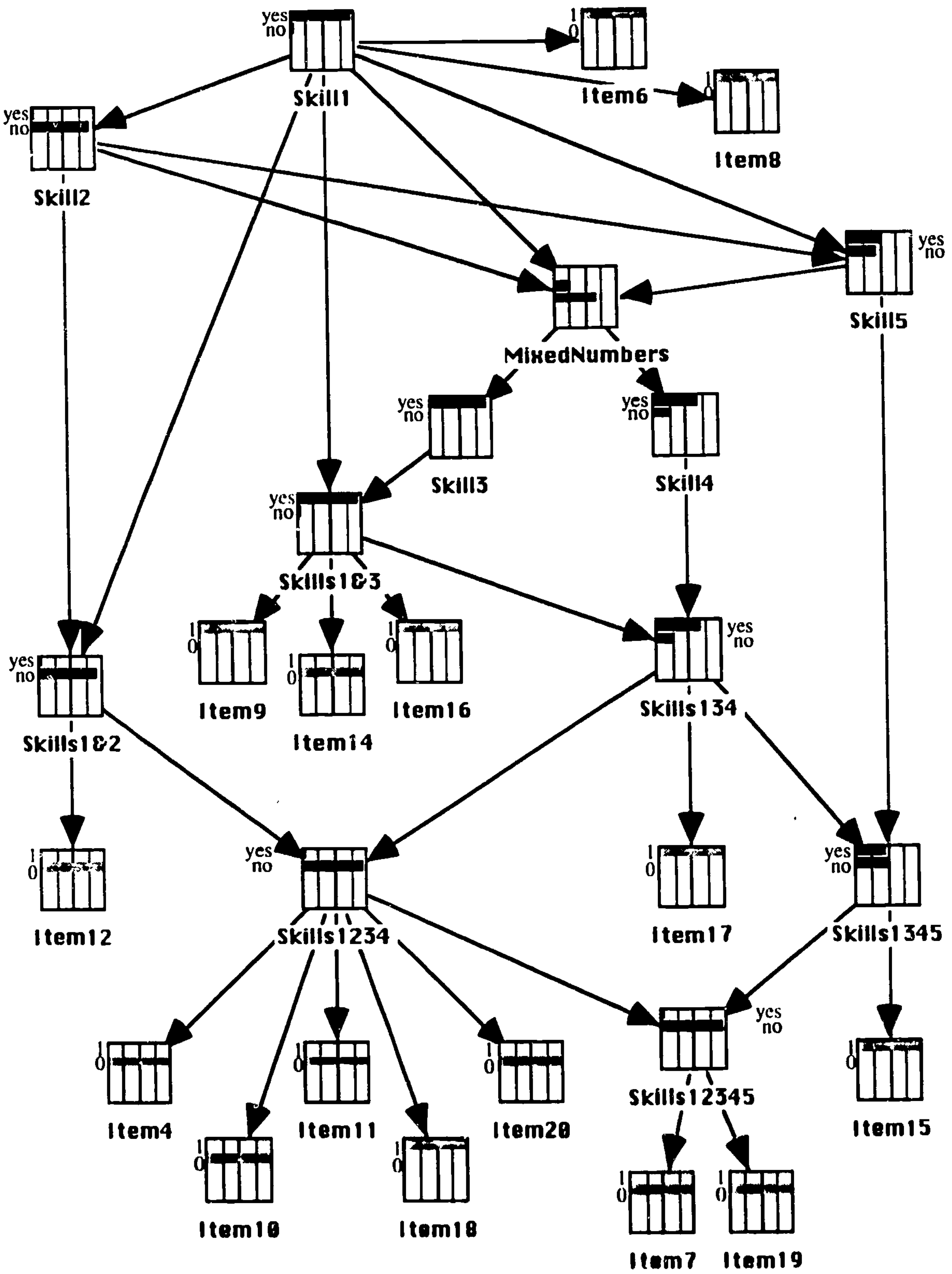
Figure 3 depicts base rate probabilities of skill possession and item percents-correct, or the state of knowledge one would have about a student we know uses Method B before observing any item responses. Figure 4 shows how beliefs change after observing mostly correct answers to items that don't require Skill 2, but incorrect answers to most of those that do. The updated probabilities for the five skills shown in Table 6 show substantial shifts away from the base-rate, toward the belief that the student commands Skills 1, 3, 4, and possibly 5, but almost certainly not Skill 2.

We built a similar network for Method A. Figure 5 incorporates it and the Method B network into a single network that is appropriate if we don't know which strategy a student uses. Each item now has three parents: minimally sufficient sets of subprocedures under Method A and under Method B, and the new node "Is the student using Method A or Method B?" An item like $7\frac{2}{3}-5\frac{1}{3}$ is hard under Method A but easy under Method B; an item like $2\frac{1}{3}-1\frac{2}{3}$ is just the opposite. A response vector with most of the first kind of items right and those of the second kind wrong shifts belief toward Method B. The opposite pattern shifts belief toward the use of Method A. A pattern with mostly wrong answers gives posterior probabilities for Method A and Method B that are about the same as the base rates, but low probabilities for possessing any of the skills. We haven't learned much about which strategy such a student is using, but we do have evidence that he isn't employing subprocedure skills effectively. Similarly, a pattern with mostly right answers again gives posterior probabilities for Method A and Method B that are about the same as the base rates, but high probabilities for possessing all of the skills. Results such as these could be used to guide instructional decisions.



Note: Bars represent probabilities, summing to one for all the possible values of a variable.

Figure 3. Prior probabilities for Method B.



Note. Bars represent probabilities, summing to one for all the possible values of a variable. A shaded bar extending the full width of a node represents certainty, due to having observed the value of that variable; i.e., a student's actual responses to tasks.

Figure 4. Posterior probabilities for Method B following item responses.

Table 6

Prior and Posterior Probabilities of Subprocedure Profile

Skill(s)	Prior probability	Posterior probability
1	.883	.999
2	.618	.056
3	.937	.995
4	.406	.702
5	.355	.561
1 & 2	.585	.056
1 & 3	.853	.994
1, 3, & 4	.392	.702
1, 2, 3, & 4	.335	.007
1, 3, 4, & 5	.223	.492
1, 2, 3, 4, & 5	.200	.003

To connect this example with the criterion-referenced testing (CRT) movement of the 1960s mentioned above, the groups of items with a common skill-set parent in Figure 2 could be viewed as a sample of tasks from a narrowly-defined behavioral domain, and probabilities of possessing the skill-set might be viewed as a tendency to perform well in that domain. The present model goes beyond the CRT framework in two ways. First, the interrelationships among such mini-domains through the delineations of procedure requirements within and across strategies provides the formerly missing connection between competence in the mini-domains and how competence develops: It develops as students learn skills and strategies that cut across mini-domains in determinable ways. Secondly, the groupings of items that are equivalent under Method A are different from the groupings based on Method B. Recognizing that the salient features of an item depend on how a student is approaching it takes a step toward addressing Thompson's (1982) question, "What can this person be thinking so that his actions make sense from his perspective?"

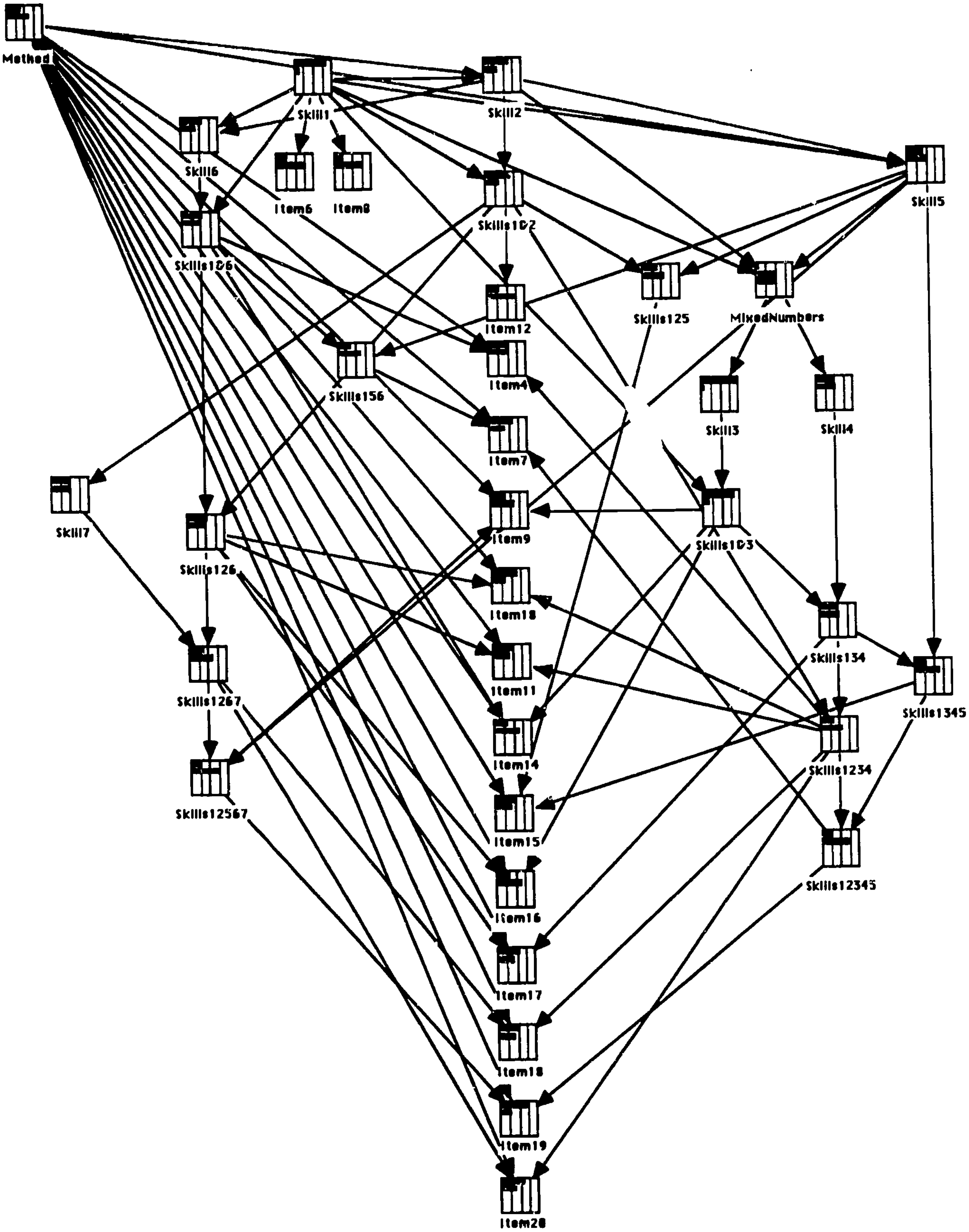


Figure 5. Prior probabilities in inference network for both methods combined.

This example could be extended in many ways, both as to the nature of the observations and the nature of the student model. With the present student model, one might explore additional sources of evidence about strategy use: monitoring response times, tracing solution steps, or simply asking the students to describe their solutions! Each has tradeoffs in terms of cost and evidential value, and each could be sensible in some applications but not others. An important extension of the student model would be to allow for strategy switching (Kyllonen, Lohman, & Snow, 1984). Adults often decide whether to use Method A or Method B for a given item only after gauging which strategy would be easier to apply. The variables in this more complex student model would express the tendencies of a student to employ different strategies under different conditions. Students would then be mixtures in and of themselves, with "always use Method A" and "always use Method B" as extreme cases. Mixture problems are notoriously hard statistical problems; carrying out inference in the context of this more ambitious student model would certainly require the richer information mentioned above. Anne Béland and I (Béland & Mislevy, 1992) tackled this problem in the domain of proportional reasoning balance-beam tasks. We modeled students in terms of neo-Piagetian developmental stages based on the availability of certain concepts that could be fashioned into strategies for different kinds of tasks. The data for inferring students' stages were their explanations of the strategies they employed on tasks with various structures.

Inference network models can play at least two important roles in educational assessment. First is the use exemplified above, short term instructional guidance, as in an intelligent tutoring system. Drew Gitomer and I are implementing probability-based inference to update the student model in an ITS for trouble-shooting an aircraft hydraulics system (Gitomer, Steinberg, & Mislevy, in press). Second is mapping the evidential structure of observations and student knowledge structures (Haertel, 1989; Haertel & Wiley, 1993). As both models and observational contexts become more complex, we must carefully sort out the informational qualities of assessment tasks to use them effectively.

Conclusion

Civil engineers designed bridges in 1893 using Euclid's geometry and Newton's laws of mechanics, in the prevailing belief that the patterns they embodied were the "true" description the universe. The variables were "the universe's" variables, with applications departing from truth only in terms of simplifications and measurement errors. The quantum and relativistic revolutions shattered this view. Yet engineers today design bridges using essentially the same formulas. Has anything changed?

The equ.^tions may be the same, but the conceptual framework within which they are comprehended is decidedly not. Today they are now viewed as engineering tools, justified to the extent that they capture patterns in nature well enough to solve the problem at hand, even as judged by the standards of the new paradigm. And while some engineers continue to attack problems that first arose in previous paradigms with a toolkit that includes methods developed under those paradigms, others attack problems that could not even be conceived last century—superconductivity, microchip design, and fusion, as examples. These problems demand a toolkit founded upon the concepts, variables, and relationships of new paradigms; some familiar tools, albeit reconceived, others totally new.

I see the analogous multiple paths of progress for educational test theory, to support inference and decision making from the perspective of contemporary psychology. Those of us in test theory must work with educators and researchers in learning areas to develop models that express key aspects of developing competence, and inferential methodologies that support defensible and cost-effective data-gathering and interpretation in practical problems. As the bridge-building analogy suggests, methodological tools developed under the trait and behavioral paradigms, properly reconceived, will serve this purposes in some applications; new tools will be needed for others. Clearly there is a lot of work to do. There are many directions to move beyond the simple psychological models and data types of familiar test theory, each presenting its own challenges. If we view ourselves as specialists in evidence and inference in school learning problems, as cast in psychological frameworks that suit those problems, clearly we can help.

References

American Council on the Training of Foreign Languages. (1989). *ACTFL proficiency guidelines.* Yonkers, NY: Author.

Andersen, S.K., Jensen, F.V., Olesen, K.G., & Jensen, F. (1989). *HUGIN: A shell for building Bayesian belief universes for expert systems* [computer program]. Aalborg, Denmark: HUGIN Expert Ltd.

Anderson, J.R., & Reiser, B.J. (1985). The LISP tutor. *Byte, 10,* 159-175.

Andrich, D. (1988, April). *A scientific revolution in social measurement.* Paper presented at the meeting of the Special Interest Group on Rasch Measurement, American Educational Research Association, New Orleans.

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-357). Hillsdale, NJ: Erlbaum.

Béland, A., & Mislevy, R.J. (1992). *Probability-based inference in a domain of proportional reasoning tasks* (ETS Research Report 92-15-ONR). Princeton, NJ: Educational Testing Service.

Bennett, R.E. (1993). Toward intelligent assessment: An integration of constructed-response testing, artificial intelligence, and model-based measurement. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 99-123). Hillsdale, NJ: Erlbaum.

Berlak, H. (1992). Toward the development of a new science of educational testing and assessment. In H. Berlak, F.M. Newmann, E. Adams, D.A. Archbald, T. Burgess, J. Raven, & T.A. Romberg, *Toward a new science of educational testing and assessment* (pp. 181-206). Albany: State University of New York Press.

Birnbaum, L. (1991). Rigor mortis: A response to Nilsson's "Logic and artificial intelligence." *Artificial Intelligence, 47,* 57-77.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443-459.

Boring, E.G. (1923). Intelligence as the tests test it. *New Republic, 34,* 35-37.

Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence, 47,* 139-159.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about the trajectories of objects. *Cognition, 9*, 117-123.

Chi, M.T.H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Clancey, W.J. (1986). Qualitative student models. *Annual Review of Computer Science, 1*, 381-450.

Cronbach, L.J., & Furby, L. (1970). How should we measure "change"—Or should we? *Psychological Bulletin, 74*, 68-80.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Deming. W.E. (1980). *Scientific methods in administration and management* (Course No. 617). Washington, DC: George Wpshington University.

Edgeworth, F.Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 599-635.

Edgeworth, F.Y. (1892). Correlated averages. *Philosophical Magazine, 5th Series, 34*, 190-204.

Estes, W.K. (1981). Intelligence and learning. In M.P. Friedman, J.P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 3-23). New York: Plenum.

Falmagne, J-C. (1989). A latent trait model via a stochastic learning theory for a knowledge space. *Psychometrika, 54*, 283-303.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.

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

Gitomer, D.H., Steinberg, L.S., & Mislevy, R.J. (in press). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 118*, 519-521.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36*, 923-936.

Glaser, R. (1991). Expertise and assessment. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition* (pp. 17-30). Englewood Cliffs, NJ: Prentice Hall.

Green, B. (1978). In defense of measurement. *American Psychologist, 33*, 664-670.

Greeno, J.G. (1976). Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), *Cognition and instruction* (pp. 123-159). Hillsdale, NJ: Erlbaum.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement test items. *Journal of Educational Measurement, 26*, 301-321.

Haertel, E.H., & Wiley, D.E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359-384). Hillsdale, NJ: Erlbaum.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: Macmillan.

Irvine, S.H., Dann, P.L., & Anderson, J.D. (in press). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology.*

Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Kirsch, I.S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults.* Princeton, NJ: National Assessment of Educational Progress/Educational Testing Service.

Krathwohl, D.R., & Payne, D.A. (1971). Defining and assessing educational objectives. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 17-45). Washington, DC: American Council on Education.

Kuhn, T.S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Kyllonen, P.C., Lohman, D.F., & Snow, R.E. (1984). Effects of aptitudes, strategy training, and test facets on spatial task performance. *Journal of Educational Psychology, 76*, 130-145.

Lauritzen, S.L., & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B, 50*, 157-224.

Lesgold, A.M., Feltovich, P.J., Glaser, R., & Wang, Y. (1981). *The acquisition of perceptual diagnostic skill in radiology* (Tech. Rep. No. PDS-1). Pittsburgh: University of Pittsburgh, Learning Research and Development Center.

Lesh, R.A., & Lamon, S. (1992). Assessing authentic mathematical performance. In R.A. Lesh & S. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 17-62). Washington, DC: American Association for the Advancement of Science.

Lesh, R., Lamon, S., Lester, F., & Behr, M. (1992). Future directions for mathematics assessment. In R.A. Lesh & S. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 379-425). Washington, DC: American Association for the Advancement of Science.

Linacre, J. M. (1989). *Multi-faceted Rasch measurement.* Chicago: MESA Press.

Lord, F.M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement, 18*, 437-454.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Martin, J.D., & VanLehn, K. (in press). OLEA: Progress toward a multi-activity, Bayesian student modeler. *Proceedings of the International Conference on Artificial Intelligence in Education.*

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mathematical Sciences Education Board (1993). *Measuring up: Prototypes for mathematics assessment.* Washington, DC: National Academy Press.

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

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments* (Research Report RR-92-39). Princeton: Educational Testing Service.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: Macmillan.

Minsky, M. (1975). A framework for representing knowledge. In P.H. Winston (Ed.), *The psychology of computer vision* (pp. 211-277). New York: McGraw-Hill.

Mislevy, R.J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement, 12*, 281-296.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R.J. (in press). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Mislevy, R.J., Yamamoto, K, & Anacker, S. (1992). Toward a test theory for assessing student understanding. In R.A. Lesh & S. Lamon (Eds.), *Assessments of authentic performance in school mathematics* (pp. 293-318). Washington, DC: American Association for the Advancement of Science.

Mislevy, R.J., Sheehan, K.M., & Wingersky, M.S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55-78.

Mosenthal, P.B., & Kirsch, I.S. (1991). Toward an explanatory model of document literacy. *Discourse Processes, 14*(2), 147-180.

Nilsson, N.J. (1991). Logic and artificial intelligence. *Artificial Intelligence, 47*, 31-56.

Noetic Systems, Inc. (1991). ERGO [computer program]. Baltimore, MD: Author.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

Push, S., & Hicks, C. (1993). Measuring performance in ways to help students learn mathematics, *National Research Council News Report, Vol. XLIII, No. 1*, 6-7

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research/Chicago: University of Chicago Press (reprint).

Resnick, L.B., & Resnick, D.P. (1989). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Conner (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer.

Rumelhart, D.A. (1980). Schemata: The building blocks of cognition. In R. Spiro, B. Bruce, & W. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Erlbaum.

Scheiblechner, H. (1972). Das lernen und lösen komplexer denkaufgaben (The learning and solution of complex cognitive tasks). *Zeitschrift für Experimentalle und Angewandte Psychologie, 19*, 476-506.

Scheuneman, J., Gerritz, K., & Embretson, S. (1991). *Effects of prose complexity on achievement test item difficulty* (Research Report RR-91-43). Princeton: Educational Testing Service.

Schum, D.A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, MD: University Press of America.

Shafer, G., & Shenoy, P. (1988). *Bayesian and belief-function propagation* (Working Paper 121). Lawrence: University of Kansas, School of Business.

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

Sheehan, K.M., & Mislevy, R.J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement, 27*, 255-272.

Shoemaker, D.M. (1975). Toward a framework for achievement testing. *Review of Educational Research, 45,* 127-147.

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan

Stake, R.E. (1991). The teacher, standardized testing, and prospects of revolution. *Phi Delta Kappan, 73,* 243-247.

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Tatsuoka, K.K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement, 24*, 233-245.

Tatsuoka, K.K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto, (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Thompson, P.W. (1982). Were lions to speak, we wouldn't understand. *Journal of Mathematical Behavior, 3*, 147-165.

VanLehn, K. (1988). *Problem solving and cognitive skill acquisition* (Tech. Rep. AIP #32). Pittsburgh, PA: Carnegie Mellon University.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions.* Cambridge, MA: MIT Press.

Vosniadou, S., & Brewer, W.F. (1987). Theories of knowledge restructuring in development. *Review of Educational Research, 57*, 51-67.

Voss, J.F., Greene, T.R., Post, T.A., & Penner, B.C. (1983). Problem-solving skill in the social sciences. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 17, pp. 165-213). New York: Academic Press.

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Educational Research* (Vol. 17, pp. 31-74). Washington, DC: American Educational Research Association.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.