DOCUMENT RESUME

ED 381 582 TM 022 054

AUTHOR Mislevy, Robert J.
TITLE Test Theory Reconceived.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; Office of Naval Research, Arlington, VA. Cognitive and Neural Sciences Div.
REPORT NO ETS-RR-94-2-ONR
PUB DATE Feb 95
CONTRACT N00014-91-J-4101
NOTE 63p.; Based on an invited address presented to the National Council of Measurement in Education (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Cognitive Psychology; Developmental Psychology; *Educational Testing; *Inferences; *Research Methodology; *Statistical Analysis; Test Interpretation; *Test Theory

ABSTRACT

Educational test theory consists of statistical and methodological tools to support inferences about examinees' knowledge, skills, and accomplishments. The evolution of test theory has been shaped by the nature of users' inferences which, until recently, have been framed almost exclusively in terms of trait and behavioral psychology. Progress in the methodology of test theory enabled users to extend the range of inference, sharpen their logic, and ground their interpretations more solidly within these psychological paradigms. In particular, the focus remained on students' overall tendency to perform in prespecified ways in prespecified domains of tasks; for example, to make correct answers to mixed-number subtraction problems. Developments in cognitive and developmental psychology broaden the range of desired inferences, especially to conjectures about the nature and acquisition of students' knowledge. Commensurately broader ranges of data-types and student models are entertained. The same underlying principles of inference that led to standard test theory can be applied to support inference in this broader universe of discourse. Familiar models and methods--sometimes extended, sometimes reinterpreted, sometimes applied to problems wholly different from those to which they were first devised--can play a useful role to this end. Contains three tables and seven figures. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 381 582

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it
Minor changes have been made to improve reproduction quality
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

RR-94-2-ONR

TEST THEORY RECONCEIVED

Robert J. Mislevy

This research was sponsored in part by the Cognitive Science Program Cognitive and Neural Sciences Division Office of Naval Research, under Contract No. N00014-91-J-4101 R&T 4421573-01

Robert J. Mislevy, Principal Investigator

Educational Testing Service
Princeton, NJ

February 1995

Reproduction in whole or in part is permitted for any purpose of the United States Government.

Approved for public release; distribution unlimited.

BEST COPY AVAILABLE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Test Theory Reconceived

5. FUNDING NUMBERS
G. N00014-91-J-4101
PE. 61153N
PR. RR 04204
TA. RR 04204-01
WU. R & T 4421573-01

6. AUTHOR(S)
Robert J. Mislevy

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

8. PERFORMING ORGANIZATION REPORT NUMBER
RR-94-2-ONR

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Cognitive Sciences
Code 1142CS
Office of Naval Research
Arlington, VA 22217-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
N/A

11. SUPPLEMENTARY NOTES
Yes

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified/Unlimited

12b. DISTRIBUTION CODE
N/A

13. ABSTRACT *(Maximum 200 words)*

Educational test theory consists of statistical and methodological tools to support inference about examinees' knowledge, skills, and accomplishments. The evolution of test theory has been shaped by the nature of users' inferences, which until recently, have been framed almost exclusively in terms of trait and behavioral psychology. Progress in the methodology of test theory enabled users to extend the range of inference, sharpen the logic, and ground their interpretations more solidly within these psychological paradigms. In particular, the focus remained on students' overall tendency to perform in prespecified ways in prespecified domains of tasks; for example, to make correct answers to mixed-number subtraction problems. Developments in cognitive and developmental psychology broaden the range of desired inferences, especially to conjectures about the nature and acquisition of students' knowledge. Commensurately broader ranges of data-types and student models are entertained. The same underlying principles of inference that led to standard test theory can be applied to support inference in this broader universe of discourse. Familiar models

14. SUBJECT TERMS	15. NUMBER OF PAGES
Bayesian inference networks, cognitive psychology, intelligent tutoring systems, item response theory, test theory.	60
	16. PRICE CODE N/A

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

and methods--sometimes extended, sometimes reinterpreted, sometimes applied to problems wholly different from those for which they were first devised--can play a useful role to this end.

Test Theory Reconceived

Robert J. Mislevy

Educational Testing Service

January, 1995

This paper is based on an invited address to the annual meeting of the National Council of Measurement in Education in Atlanta, April, 1993. I am grateful to the organizer of the session, Suzanne Lane, and to the discussants, Robert Glaser, H.D. Hoover, and Richard Snow. Their comments have been incorporated, as have those of Isaac Bejar, Kalle Gerritz, Ivo Molenaar, Howard Wainer, and Rebecca Zwick. The work was supported by (1) Contract No. N00014-91-J-4101, R&T 4421573-01, from the Cognitive Science Program, Cognitive and Neural Sciences Division, Office of Naval Research, (2) the National Center for Research on Evaluation, Standards, Student Testing (CRESST), Educational Research and Development Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education, and (3) the Statistical and Psychometric Research Division of Educational Testing Service. The HYDRIVE example is based on a project by Armstrong Laboratories of the United States Air Force and directed by Drew Gitomer.

Copyright © 1995. Educational Testing Service. All rights reserved

Abstract

Educational test theory consists of statistical and methodological tools to support inference about examinees' knowledge, skills, and accomplishments. The evolution of test theory has been shaped by the nature of users' inferences, which, until recently, have been framed almost exclusively in terms of trait and behavioral psychology. Progress in the methodology of test theory enabled users to extend the range of inference, sharpen the logic, and ground their interpretations more solidly within these psychological paradigms. In particular, the focus remained on students' overall tendency to perform in prespecified ways in prespecified domains of tasks; for example, to make correct answers to mixed-number subtraction problems. Developments in cognitive and developmental psychology broaden the range of desired inferences, especially to conjectures about the nature and acquisition of students' knowledge. Commensurately broader ranges of data-types and student models are entertained. The same underlying principles of inference that led to standard test theory can be applied to support inference in this broader universe of discourse. Familiar models and methods—sometimes extended, sometimes reinterpreted, sometimes applied to problems wholly different from those for which they were first devised—can play a useful role to this end.

Keywords: Bayesian inference networks, cognitive psychology, intelligent tutoring systems, item response theory, test theory.

Summary test scores, and factors based on them, have often been thought of as "signs" indicating the presence of underlying, 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.

Introduction

Test theory, as it is usually thought of, 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 admixture of ideas from measurement, education, and psychology. This framework generates a universe of discourse: the nature of the educational problems and potential solutions one defines, the purposes and values of assessment, the kinds of statements one makes about students, and the ways one gathers data to inform and support these statements. Test theory, as it is usually thought of, is machinery for addressing inferential problems within this framework: What kinds of evidence are needed to support inferences about students? How much faith can be placed in the evidence, and in the ensuing statements? Are elements of evidence overlapping, redundant, or contradictory? When must different questions be asked or additional situations posed to distinguish among competing explanations of what is observed?

The emerging paradigm of cognitive psychology, with its focus on the nature and the acquisition of competence, prompts new considerations about how to collect and interpret evidence about students' learning. This paper argues that aspects of the models and methods that have evolved within standard test theory can be extended, augmented, and reconceived to address problems cast in this broader universe of discourse. The argument can be summarized as follows:

- The methodological paradigm referred to here as standard test theory arose under the psychological paradigm of trait psychology. The target of inference is a person's tendency to act in prespecified ways in prespecified domains of situations (e.g., to make correct rather than incorrect answers to multiple-choice test items).

- This methodological approach was readily adapted and extended to support assessment cast in terms of behavioral psychology, addressing domains of tasks which modifying behavioral tendencies toward was the goal of instruction (e.g., increasing students' chances of writing coherent essays on specified topics). The nature of competence so defined, and the processes by which competence increases, lay largely outside the universe of discourse of test theory proper.

- The cognitive and developmental psychological paradigms extend inquiry to the nature and the acquisition of knowledge and skills. Inferences cast in these terms encounter issues of weight and coverage of evidence, just as do inferences cast in terms of trait and behavioral psychology. The same inferential principles that led to standard test theory can be gainfully applied to this end—sometimes even some of the same models and methods, albeit construed from the perspective the operative psychological paradigm.

To accomplish this objective, it is necessary to disentangle the statistics from the psychology in standard test theory; to view test theory as the application of more general principles of inference (Mislevy, 1994). General issues of evidence in inference (see Schum, 1987, 1994), including the role of paradigms in scientific and practical work, are first discussed. The interplay between methodological and psychological paradigms in educational assessment is then addressed, with emphasis on considerations prompted by cognitive and developmental psychology. Examples from current projects illustrate central points.

Evidence, Inference, and Paradigms

Inference is reasoning from what one knows and what one observes, to explanations, conclusions, or predictions. One attempts to establish the weight and coverage of evidence in what is observed. The very first question that must be addressed is "Evidence about what?" Schum (1994, p. 20) points out the crucial distinction between *data* and *evidence:* "A datum becomes evidence in a particular inference when its relevance to this inference has been established." The same observation can be direct evidence for some conjectures and indirect evidence for others, and wholly irrelevant to still others. In

educational assessment, one observes specific actions or products that students produce in specific circumstances, sometimes as interpreted by specific observers. These are the data. To evaluate progress or guide further instruction, however, one talks at a higher level of abstraction, using specific observations as evidence for subsequent inferences. To use Kuhn's (1970) term, these more abstract conjectures must be constructed within some "paradigm" for the nature and the acquisition of competence.

Kuhn used this term to describe a set of interrelated concepts that frames research in a scientific field. Of all the phenomena that can be experienced directly or indirectly, a paradigm focuses on patterns in a circumscribed domain. The patterns, and the language and the concepts used to express them, determine the kinds of things that will be talked about and the particular things that can be said. A paradigm frames what is construed as problems, and how attempts to solve them are to be evaluated. Most scientific research is carried out within an existing paradigm. Kuhn referred to solving the outstanding puzzles a paradigm poses as "normal science"—improving measurements, developing inferential machinery, working out relationships in greater detail, extending ideas to new situations, and integrating 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.

"Scientific revolutions," in which a new major paradigm displaces an existing paradigm, were Kuhn's focus. 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.

As an example, 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 equations may be the same, but the conceptual framework within which they are comprehended is decidedly not. Today the equations are 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 quantum and relativistic paradigms. (Bridges are neither so small as too require quantum models nor so fast-moving as to demand relativistic corrections.) And while some engineers continue to attack problems such as bridge-building that first arose under previous paradigms with a toolkit containing many methods developed under those paradigms, other engineers attack problems that could not even be conceived last century—superconductivity, microchip design, and fusion, to name a few. These problems demand a toolkit founded upon the concepts, variables, and relationships of new paradigms; some familiar tools, albeit reconceived, others totally new.

Psychological Paradigms and Test Theory

A conception of student competence and a purpose for assessment determine the kind of information that is needed for an assessment, and should drive in turn the particular methods that are needed to get students to act in ways that reveal something about their competencies—that is, the forms of assessment (Berlak, 1992). The following sections discuss implications that the trait, behaviorist, and cognitive psychological paradigms hold for conceptions of competence. It is beyond the scope of this presentation to consider all the ways that different purposes entail different evidential requirements, even under a given conception of competence; the reader is referred to Millman and Greene (1989) on dimensions of purpose that shape the form of assessments.

As noted above, test theory is machinery for reasoning from students' behavior to conjectures about their competence, as framed in a particular conception of competence. In any particular application, this conception takes the form of a set of aspects of skill and knowledge that are important for the job at hand, whether that job be summarizing the competencies students have acquired thus far or guiding instruction to increase their competencies further. These are the variables in what might be called a "student model"—a simplified description of selected aspects of the infinite varieties of skills and knowledge

that characterize real students. Although this is the level at which one evaluates students' learning and plans further instruction, these variables are not directly observable.

Depending on the purpose, one might distinguish from one to hundreds of aspects of competence in a student model. 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. Depending on the purpose of the assessment and the operative psychological paradigm, they might concern tendencies in behavior, conceptions of phenomena, available strategies, or levels of development. At one extreme, "verbal and quantitative ability" are the only two variables in the student model underlying the Scholastic Assessment Test. At the other extreme, the student model in John Anderson's LISP tutor (Anderson & Reiser, 1985) concerns mastery of hundreds of production rules (some correct, others erroneous) in sufficient detail to provide answers to any problem in a task domain.

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." Hypothetical (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. Under trait psychology, the variables in the student model are the values of the traits of interest.

When Charles Spearman used scores on a fixed set of knowledge and puzzle-solving tasks to "measure intelligence," the notion of a trait was not new. Paul Broca had attempted to assess "intelligence" in the previous century by charting cranial volumes, as had Francis Galton by measuring reaction times. Neither was the idea of observing behavior in samples of standardized situations new. Three thousand years earlier, the Chinese discovered that observing an individual's performance under controlled conditions could support 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 so obtained were accepted 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 conjoining of a psychological and a methodological paradigm suited the mass educational system that also arose in the United States at the turn of the century (Glaser, 1981). Educators saw 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 number, broadly construed, stable over time, applicable to wide ranges of students, and discernible with data that were easy to gather and interpret.

Basic concepts in 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 methodological paradigm of characterizing people's tendencies to behave in prescribed ways in prescribed settings. The inferential considerations that motivated these developments merit a brief review because they transcend the substantive content of the psychological paradigm under which test theory arose.

Edgeworth (1888, 1892) and Spearman (1904, 1907) launched true-score, or classical test theory (CTT) by applying mathematical models and statistical tools from physical measurement to what were seen as 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," or the hypothetical expected response across the entire domain of tasks. 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 similarly-structured 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 weight and the same scope of coverage, about the same true score.

An inferential concept that plays a central role in test theory is *conditional independence*. In statistical terms, variables may be related in a population, but they are conditionally independent if they are unrelated given the values of another set of variables. The importance of conditional independence in reasoning is that it expresses the explanatory relationships and generative principles of a substantive paradigm, around which inference within that paradigm can be structured:

> *[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

In CTT, interest centers on the unobservable variable "true score;" this is the student model, expressing the aspect of knowledge and skill in terms of which inferences will be based. Observable scores on actual parallel tests are posited to be conditionally independent given true score (Lord & Novick, 1968). Spearman's *methodological* insight (as distinguished from his thoughts about human abilities *per se,* or his student model) was this: Conditional independence of observable test scores, given an unobservable "intelligence" variable, implies particular patterns of relationships among the observable scores. This insight provides a framework for organizing observations, for quantifying the evidence about true scores provided by observed scores, and, at least in principle, for disconfirming conjectures about behavior in terms of true scores and hence of hypothesized traits. Test theorists have since been working out and extending the logic of inference in terms of unobservable variables—exploring the possibilities and the limitations, developing statistical machinery for estimation and prediction (Lewis, 1986).

The indicator of a test's evidential value under CTT was *reliability*, the correlation between parallel forms in a specified population of examinees (which can be estimated from actual parallel forms when they are available, or, when tests consist of sets of exchangeable

tasks, from the internal consistency of tasks within a single test). 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 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 evidential value of test scores for other inferences, 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.

Applying 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 have come to be applied. One hopes 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, perhaps as defined 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 thus extends the application of CTT machinery for inference concerning domain proficiency beyond selection and placement decisions framed in terms of trait psychology, to instructional planning and evaluation problems framed in terms 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 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.*
>
> D.R. Krathwohl & D.A. Payne, 1971, p. 17-18.

A 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 in this way used to support instructional problems cast in behavioral psychology.

Extending the methodology

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, and so on. Generalizability theory greatly expands the range of conjectures one can address—though still within a universe of discourse in which inferences concern "overall tendency toward specified behavior in a specified domain," as defined from the point of view of the test designer.

Item response theory (IRT; Hambleton, 1989) originated in the early 1940's 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 prespecified domain of tasks. Beyond CTT, IRT posits a conditional independence relationship among individual test items given examinees' proficiency variables. This conceptualization helps solve some practical problems that could be expressed in the overall-proficiency paradigm, but were poorly handled with CTT tools (e.g., characterizing the accuracy of estimation for individual examinees, constructing tests with desired properties, and tailoring tests to examinees in response to their continuing sequence of responses). Focus remains on overall proficiency, but regularities in relationships between this overall proficiency and behavior on specific tasks are exploited.

Rapid progress has been made by applying developments in statistics to IRT (e.g., Bock & Aitkin, 1981; Lord, 1980; Mislevy, 1991).

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—more than merely improving applications within the paradigm, 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 phenomenon" (Kuhn, 1970, p. 205). Example 1 below remarks on diagnostic tools for using IRT in light of results from cognitive psychology.

Another 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 joint distributions of scores of several tests, to the end of better understanding the meaning of variables they define. A researcher might seek recurring patterns in tests with systematically varying tasks; for example, a 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 within the CTI paradigm 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" (Messick, 1989, pp. 13-14).

At its leading edge, if not in everyday practice, test theory for mental measurement has come of age—in the sense of having developed methodological tools for gathering and interpreting data, and a 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

Through the use of standard test theory, evidence can be characterized and brought to bear on inferences about students' overall proficiency in behavioral domains, for determining a students' levels of proficiency, comparing them to others or to a standard, or gauging changes 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 overarching 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 (1970) caution that the characterization of change may lay beyond reach the reach of familiar test theory:

> *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)

Cognitive Psychology

In contrast to schooling applications, most contemporary research into human abilities does not take place within the trait or behavioral psychological paradigms, but within what has come to be called the cognitive paradigm. Three key propositions from cognitive psychology (paraphrasing Lesh & Lamon, 1992, p. 60) hold implications for instruction and assessment:

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.

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. While experts in various fields of learning do generally command more facts and concepts than novices, and have richer interconnections among them, the real 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):

> *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.*
>
> Rumelhart, 1980, p. 55.

A schema is "instantiated" when one perceives some of its relationships in a situation, which focuses attention on filling in missing variables, inferring additional relationships, and checking for specifics at odds with usual expectations. Much of this activity is unconscious and automatic, as when one perceives letters in the course of reading a text. Sometimes aspects of it are conscious and deliberate, as when trying 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; people must ever and always construct meaning, in terms of knowledge structures that 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).

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 in fact merely adding bits to existing structures. Sometimes it involves generalizing or connecting schemas. Other times it involves wholesale abandonment of important parts of schemas, with replacement by qualitatively different structures (Rumelhart, 1980).

Less is known about actual mechanisms underlying these changes than about conditions that seem to facilitate them: 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 how case studies are used in fields where the goals of instruction concern 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. Developing expertise is generally characterized by the increase of so-called metacognitive skills: self-awareness of using models, and acquiring skill and flexibility in how to construct them, modify them, and adapt them to the problem at hand (Glaser, Lesgold, & Lajoie, 1987).

Implications for assessment

> *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]
>
> R. Glaser, 1991, p. 26.

The first questions in any assessment should be, "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 that the educational experiences are meant 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 assessment to meet all the objectives this analysis might yield; only models more or less useful for various purposes, by virtue of the patterns among observations they can express and the information they can convey thereby. There is no single "best" method for gathering data; only methods more or less effective at evoking evidence for the inferences to be made. These factors can vary dramatically across applications. The following issues are confronted whenever one attempts to frame assessment within a cognitive paradigm.

The nature of the "student model." Obviously any student model oversimplifies the reality of cognition—whatever that may be! In real-world educational assessment, as Greeno (1976, p. 133) points out, "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." For immediate feedback for short-term instructional decisions, as in intelligent tutoring systems, there is a need for more detail in the student model. For example, John Anderson's LISP tutor characterizes programming competence in terms of a specified set of hundreds of production rules, or condition-action relationships (Anderson & Reiser, 1985).

For accountability purposes, on the other hand, a coarser grain-size may well suffice. Gathering detail with no intended use would waste scarce resources, if only summary indicators of learning are required to monitor progress—assuming these indicators are supplemented by more detailed and more focused assessment to guide instruction and curricular refinements along lines that are consistent with the indicators. For example, Table 1 shows excerpts from the American Council of Teachers of Foreign Language (ACTFL) guidelines for reading proficiency (ACTFL, 1989). Assessors map students' observed behavior into this abstract frame of reference, based on theories and observations of second language acquisition. 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. Example 2 below concerns the problem of making abstractly stated guidelines meaningful in practice.

[Table 1 about here]

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. Standard mental measurement paradigm attends to the problem stimulus strictly from the assessor'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 [the] products of the understanding process, rather than the problem stimulus itself" (VanLehn, 1989, p. 532). Because different knowledge structures can lead to the same behavior, observed behavior constitutes *indirect* evidence about cognitive structure—which can be crucial in an application such as tutoring (see, e.g., Frederiksen & White, 1988). 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.

Compared with inference about behavioral tendencies, a chain of inference that ends with conjectures about knowledge structures has additional links, additional sources of uncertainty. Forging this chain requires knowing how competence in the domain develops. The inferential challenges routinely faced under the standard mental measurement paradigm, such as limited information and multiple sources of uncertainty, do not disappear when interest shifts to 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, in order to frame conjectures and interpret observations in their light. What are the important concepts and relationships students are to learn, and how do they learn them? What evidence can be obtained to gauge their progress, and help determine what they should do next? Disambiguating alternative explanations of behavior may also require having to gather not merely more evidence, but of complementary sources of evidence (Martin & VanLehn, 1993).

Effective assessment under a cognitive perspective requires, first and foremost, being clear about exactly what inferences one wants to make. This done, strategies and techniques analogous to those long used to make inference under the mental measurement paradigm more efficient: Avoiding the collection of *data* that hold little value as *evidence* for the targeted inferences. 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 evidence 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).

The role of conditionality in inference. While test scores do reveal 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 trait/behavioral perspective, this is "noise," or measurement error, leading to low reliability under CTT, low generalizability under generalizability theory, and low item discrimination parameters under IRT. It obscures what one is interested in from that perspective, namely, locating people along a single dimension as to a *general* behavioral tendency. For inferences concerning overall proficiency, tasks that don't line up people in the same way are less informative than ones that do.

Such interactions are fully expected from the cognitive perspective, however, since knowledge typically develops first in context, then is extended and decontextualized so it can be applied to more broadly to other contexts. The in-depth project that provides solid information about students whose prior knowledge structures it dovetails, becomes an unconscionable waste of time for students for whom it has no connection. The same task can therefore reveal either vital evidence or little at all, depending on the target of inference and the relationship of the information it carries to what is known from other sources. How to deal with these interactions in assessment depends largely on the purpose of an assessment.

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;* most students would perform poorly in most contexts, and only a few contexts—and different ones for different students—would provide clues to their nascent understanding. 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 (such as that provided the ACTFL Reading Guidelines), 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.

Examples

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

As Snow and Lohman (*op cit.*) note, sometimes it really is useful to know how proficient students are in certain domains of tasks, as indicated by on their performance on a sample of those tasks. But while the trait and behavioral paradigms end with statements about tendencies in the behaviors of interest, a cognitive perspective can offer benefits of several kinds even when standard test theory is used to gather, summarize, and characterize evidence in these applications; for example, in (1) defining and structuring the domain of tasks, (2) enriching the interpretation of scores, and (3) identifying students for whom the single-number score is misleading. This section illustrates some these ideas in a measure of document literacy.

The sixty-three tasks comprising the Survey of Young Adult Literacy (SYAL; Kirsch & Jungeblut, 1986) were designed to evoke the skills people need to locate and use information contained in non-prose formats such as forms, tables, charts, signs, and catalogs. Most required open-ended responses. 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 item response theory (IRT) model was fit, and distributions of overall proficiency in terms of an IRT variable were produced. An 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. Under the Rasch (1960) model for dichotomous items,

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

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:

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

The IRT model 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 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. Creating the domain of the tasks was the crucial step in establishing the overall proficiency which would constitute the operational definition of document literacy. 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 noted above, these match-ups vary from one person to another for any given task. . IRT item difficulty parameter captures only 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 will therefore forego potential information in any given person's responses 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 (e.g., Sternberg, 1977). 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-percent 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 (1973) 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. Incorporating a residual term to allow for variation of difficulties among items with the same key features gives

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

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 variance ϕ^2.

Sheehan and Mislevy (1990) fit this model with item features based on Mosenthal and Kirsch's (1988) cognitive analysis of the difficulty of document literacy tasks. Mosenthal and Kirsch first characterize the information contained in documents and document task directives according to three levels of organization: (1) the organizing category, (2) the specific category, and (3) the semantic feature. Semantic features are bits of information that belong to specific categories, which are nested within distinct organizing categories. After parsing materials and directives in these terms, they defined three classes of variables they expected to correlate with task difficulty: (1) variables that characterize the length and organizational complexity of the *materials* which document tasks refer to; (2) variables that characterize the length and organizational complexity of task *directives*; and (3) variables that characterize the difficulty of the task solution *process*. 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 can help item writers control the difficulty of the tasks they develop (see, for example, Embretson, 1985). 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.

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. An examinee with θ=1, for example, might be expected to manage

unfamiliar tasks that require matching information across two organizing categories, but have only even odds on tasks with requiring three matches. While the IRT θ still only captures overall competence, this connection adds a layer of meaning to score interpretation.

The same connection offers practical benefits: Exploiting information from the cognitive perspective can reduce or even eliminate pretesting meant to estimate item parameters (Mislevy, Sheehan, & Wingersky, 1993), thus opening the door to using IRT with tasks created in real time with generative algorithms based on cognitive processing models (Bejar, 1993; Irvine, Dann, & Anderson, in press). It must be recalled, of course, that all this modeling is just "on the average." 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 (e.g., Smith, 1986; Tatsuoka, 1990). The same processing-feature structure can be used to examine unexpected response patterns of individual respondents, complementing overall-proficiency θ estimates with diagnostic information.

Example 2: AP Studio Art Portfolios

Performance assessment commands attention partly because it provides direct evidence about productive aspects of knowledge, and partly because of its potential positive impact on educational practice (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. That tasks stimulate creative or problem-solving thinking is to no avail unless the critical information for the targeted inferences can be distilled from the performance. This example is based on Myford and Mislevy's (1995) study of the 1992 College Board's Advanced Placement (AP) Studio Art portfolio assessment, which addresses the issues of establishing shared standards for recognizing what is important in performance and mapping it into a summarizing structure. It illustrates how the machinery of IRT, created to model regularities in observed behavior under the presumption that student by task interactions were "noise," can be used to model regularities in judges' ratings under the presumption that student by task interactions are expected and to be "conditioned out" of the rating process.

The AP Studio Art portfolio assessment includes 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 requirements set forth in the AP Studio Art materials. The portfolio requirements are intended to ensure that evidence about key aspects of artistic development will be evoked.[1] For example, Section A consists of four works submitted in original form to be rated as to "overall quality," and Section B, the student's "concentration," consists of up to 20 slides, a film, or a videotape illustrating a concentration on a student-selected theme and a paragraph or two describing the student's goals, intentions, influences, and other factors that help explain the series of works. 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.

The Myford and Mislevy study uses two distinct perspectives, "statistical" and "naturalistic," which are required in tandem to analyze and improve a system the size of AP

---

[1] 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. 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. 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, Behr, & Lester, 1992, p. 407).

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 thinking about quality control in industry (e.g., Deming, 1982). One begins by establishing a statistical framework for analyzing data, to quantify typical and expected sources of variation (in this 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 this 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.

The statistical component employs Linacre's (1989) FACETS model, an extension of the Rasch model of the previous example. In this application, 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 Portfolio Section h with K ordered scale categories is given as

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

where ξ_k is the "harshness" parameter associated with Judge j, η_h is a "section difficulty" parameter, 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. This model applies the regularity patterns embodied in IRT beyond the original "tendency for specified behavior on specified tasks" usage; the same mathematical structures address regularities in readers' application of common standards to possibly quite different behaviors in different contexts. For example, 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." The pertinent question is not how well the student who painted landscapes would have fared with angularity in ceramics, but how consistently raters viewing either concentration would map the performances into the same evaluative framework. 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 the course is intended to promote, namely, conceptualizing and confronting one's own challenges.

In essence, FACETS fits a main-effects model to log-odds of ratings. Variation among students, as a main effect, is anticipated. Estimates of portfolio "measures" can be obtained, along with corresponding indices of the degree of uncertainty associated with those estimates—that is, a characterization of the weight of evidence, taking into account such factors as variation among readers and among sections, and the extent of disagreement among 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. Little variation of this type was present 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 may reduced by such means as improving reader training, sharpening the definition of standards, or distinguishing aspects that should be rated separately. Particular reader/portfolio combinations that are especially unusual in view of the main effects are highlighted in the form of outliers from the model.

By identifying outliers, statistical analyses can indicate where to focus attention—but not 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 the project. Discussion were held with experienced readers of 9 portfolios each for Section A and Section B that received highly discrepant ratings, in order to gain insights into the judging process in general, and into the features that made rating these particular portfolios difficult. Avenues for exploration that emerged in these discussions included 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. This study illustrates how models originally developed under the trait psychological paradigm but extended to a cognitive/developmental paradigm can be employed to characterize the weight of evidence about target inferences, and to provide information to monitor and improve the system over time.

Example 3: Mixed Number Subtraction

The form of the data in this 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 (Mislevy, 1995). 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, the 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.

The responses of 530 students to 15 items were analyzed. As shown in Table 2, each item was characterized in terms of which of seven subprocedures required to solve it with Method A and those 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 the student 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, Tatsuoka, 1990, and Haertel & Wiley, 1993). However, sometimes students miss items even under these conditions (false negatives), and 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.

[Table 2 about here]

Inference in complex networks of interdependent variables such as these 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; see Béland & Mislevy, 1992, Martin & VanLehn, 1993, and Mislevy, 1995, for further discussion of inference networks in cognitive assessment).[2] Inference networks exploit conditional independence relationships—in this example, conditional independence of item responses given procedure knowledge *and* strategy usage. Figure 1 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 "Skills1&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 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

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

from Tatsuoka's (1983) "rule space" analysis of the data, with only students classified as Method B users.

[Figure 1 about here]

Figure 2 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 3 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 3 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.

[Figures 2 & 3 and Table 3 about here]

A similar network was built for Method A. Figure 4 incorporates it and the Method B network into a single network that is appropriate if one doesn't know which strategy a student uses. Each item now has three parents: minimaliy 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. One learns little about which strategy such a student is using, but there is evidence that subprocedure skills are not being employed 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.

[Figure 4 about here]

To connect this example with the criterion-referenced testing (CRT) movement of the 1960's mentioned above, the groups of items with a common skill-set parent in Figure 1 could be viewed as a sample of tasks from a narrowly-defined behavioral domain, and probabilities of the 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 delineation 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 toward addressing Thompson's (1982) question, "What can this person be thinking so that his actions make sense from his perspective?"

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. Béland and Mislevy (1992) tackled this problem in the domain of proportional reasoning by additionally using students' explanations of solutions of balance-beam problems.

Example 4: An Intelligent Tutoring System

Intelligent tutoring systems (ITSs) are predicated on some form of student modeling to guide tutor behavior. Inferences about what a student knows and does not know can affect the presentation and pacing of problems, quality of feedback and instruction, and determination of when a student has completed some set of tutorial objectives. This example discusses the HYDRIVE ITS (Gitomer, Steinberg, & Mislevy, 1995), which, in the course of implementing principles of cognitive diagnosis, adapts a number of test theory concepts and tools to implement principles of probability-based reasoning.

HYDRIVE is an intelligent video-disc based tutoring/assessment system designed to facilitate the development of troubleshooting skills for the F-15 aircraft's hydraulics systems. Hydraulics systems are involved in the operation of flight controls, landing gear, the canopy, the jet fuel starter, and aerial refueling. HYDRIVE is designed to simulate many of the important cognitive and contextual features of troubleshooting on the flight line. A problem begins with a video sequence in which a pilot, who is about to take off or has just landed, describes some aircraft malfunction to the hydraulics technician (e.g., the rudders do not move during pre-flight checks). HYDRIVE's interface then offers the student several options, such as the following: performing troubleshooting procedures by accessing video images of aircraft components and acting on those components; reviewing on-line technical support materials, including hierarchically organized schematic diagrams; and making an instructional selections at any time during troubleshooting, in addition to or in place of instruction the system itself recommends. The state of the aircraft system, including changes brought about by user actions, is modeled by HYDRIVE's system model. Performance is monitored by evaluating how the student uses available information, as chronicled in the system model, to direct troubleshooting actions. HYDRIVE's student model is used to diagnose the quality of specific troubleshooting actions, and to characterize student understanding in terms of more general constructs such as knowledge of systems, strategies, and procedures that are associated with troubleshooting proficiency.

The grain-size and the nature of a student model in an ITS ought to be targeted to the instructional options available (Kieras, 1988). A model will first need to include a set of cognitive features related to performance, as revealed by analyses of the skills and understandings needed for accomplished performance. But because accomplished performance derives from the complex structuring of knowledge and skills, a cognitive model of student performance in an ITS will thus need to represent the interrelationships of target skills and understandings.

Wenger (1987) describes three levels of information that such a model might address. Early ITSs focusing on the *behavioral level* were usually concerned with the correctness of student behaviors referenced against some model of expert performance. For example, SOPHIE-I (Brown, Burton & Bell, 1975) contrasted student behaviors with domain performance simulations as a basis for offering corrective feedback. The *epistemic level* of information is concerned with particular knowledge states of individuals. The SHERLOCK ITS (Lesgold, Eggan, Katz, & Rao, 1992) makes inferences about the goals

and plans students are using to guide their actions during problem solving. Feedback is meant to respond to "what the student is thinking." The *individual level* addresses broader assertions about the individual that transcend particular problem states. Whereas the epistemic level of diagnosis might lead to the inference that "the student has a faulty plan for procedure X", the individual level of information might include the assertion that "the student is poor at planning in contexts A and B."

This individual level of information has received the least attention in the field of intelligent tutoring assessment. Conversely, test theory has focused mainly on the individual level, with little explicit attention to the epistemic level. One might assert, for example, that an individual has high ability in mathematics, without an account of the epistemic conditions that characterize high ability. By bridging between the individual and epistemic levels of information, a student model can have both the specificity to facilitate immediate feedback in a problem-solving situation, and the generality of individual information to help sequence problems, moderate instruction, and track proficiency in general terms. The HYDRIVE student model is designed to support generalized claims about aspects of student troubleshooting proficiency from detailed epistemic analysis of particular actions within the system. Abstractions, such as a student's *strategic understanding*, are the target constructs of the troubleshooting domain upon which the instructional components focus.

Figure 5 is a simplified version of portions of the inference network through which the HYDRIVE student model is operationalized and updated within a given problem. Four groups of variables can be distinguished: (1) The rightmost nodes are the "observable variables," actually the results of rule-driven epistemic analyses of student's actions in a given situation. (2) Their immediate parents are knowledge and strategy requirements for two prototypical situations addressed in this simplified diagram. (3) The long column of variables in the middle concerns aspects of subsystem and strategic knowledge, corresponding to instructional options. (4) To their left are summary characterizations of more generally construed proficiencies. The structure of the network, the variables that capture the progression from novice to expert hydraulics troubleshooter, and the conditional probabilities implemented in the network are based on in-depth analyses of experts and novices verbalizations of their problem-solving actions (Means & Gott, 1988), and the observation of small numbers of students actually working through the problems in the HYDRIVE context.

[Figure 5 about here]

Strictly speaking, the values of the *"observable" variables* are not observable behaviors, but interpreted outcomes of analyses of episodes of students' actions that characterize a sequence as, for example, inferred "space-splitting" in situations in which this is possible, "serial elimination," "redundant action," "irrelevant action," or "remove-and-replace"—all defined in terms of the current state of the system model. HYDRIVE employs a relatively small number of interpretation rules (~25) to classify each troubleshooting action in terms of both the student and the best strategy. An example of a student strategy rule is:

> *IF active path which includes failure has* **not** *been created and the student creates an active path which does* **not** *include failure and edges removed from the active problem area are of one power class, THEN the student strategy is power path splitting.*

As potential observable variables, these action episodes are not predetermined and uniquely-defined in the manner of usual assessment items, since a student could follow a virtually infinite number of paths through the problem. Rather than attempting to model all possible system states and specific possible actions within them, HYDRIVE posits equivalence classes of states, each of which could arise many times or not at all in a given student's work. Members of these equivalence classes are treated as conditionally independent, given the status of the requisite skill and knowledge requirements. Two such classes are illustrated in Figure 5: A canopy situation in which space-splitting is not possible, and a landing gear situation in which space-splitting is possible. Figure 5 depicts belief after observing, in three separate situations from the canopy/no-split class, one redundant and one irrelevant action (both ineffectual troubleshooting moves) and one remove-and-replace (serviceable but inefficient). Serial elimination would have been the best strategy in this case, and is most likely when the student has strong knowledge of this strategy and all relevant subsystems. Remove-and-replace is more likely when a student possesses some subsystem knowledge but lacks familiarity with serial elimination. Weak subsystem knowledge increases chances of irrelevant and redundant actions. All interpreted actions are possible from all combinations of student variable values; sometimes students with good understanding carry out redundant tests, for example, and sometimes students who lack understanding unwittingly make the same action an expert would. These possibilities must be reflected in the conditional probabilities of actions, given the values of student model variables.

Subsystem and strategy variables are meant to summarize tendencies in interpreted behaviors at a level addressed by instruction, and to disambiguate patterns of actions in light of the fact that inexpert actions can have several causes. As a result of the three inexpert canopy actions discussed above, Figure 5 shows belief shifted toward lower values for serial elimination, and for all subsystem variables directly involved in the situation (mechanical, hydraulic, and canopy knowledge). Any or all could be a problem, since all are required for high likelihoods for expert actions. Variables for subsystems not directly involved in these situations are also lower, because to varying extents, students familiar with one subsystem tend to be familiar with others, and, to a lesser extent, students familiar with subsystems tend to be familiar with troubleshooting strategies. These relationships are expressed by means of the more *generalized system and strategy knowledge* variables at the left of the figure. These variables serve to exploit the indirect information about aspects of knowledge not directly tapped, and to summarize broadly construed aspects of proficiency for evaluation and problem-selection.

Figures 6 and 7 represent the state of belief that would result after observing two different sets of actions in situations involving the landing gear, in which space-splitting is possible. Figure 6 shows the results of three more inexpert action sequences. Status on all subsystem and strategy variables is further downgraded, and reflected in the more generalized summary variables. Figure 7 shows the results of observing three good actions: two space-splits and one serial elimination. Belief about strategic skill has increased, as have beliefs about subsystems involved in the landing gear situations. Problems in mechanical and canopy subsystem knowledge are now the most plausible explanations of the three inexpert canopy situation actions. The diffuse belief at the generalized proficiency level results from the uneven profile of subsystem knowledge, despite fairly accurate information about individual aspects of the student's knowledge.

[Figures 6 and 7 about here]

Large numbers of solutions from acknowledged experts and novices of various types were not available to estimate the conditional probabilities in the HYDRIVE inference network. Initial values were set subjectively, and revised through an iterative model-checking process: positing values, entering actions suggested from the cognitive task analysis as proxies for what the student might do within the tutor, then evaluating the behavior of the network to determine whether all nodes were behaving sensibly in terms of the cognitive model. The initial probabilities were unsatisfactory in several ways. At

times, student estimates would be updated too rapidly. At other times, they wouldn't be updated despite actions that should affect belief about student competence. Some updates moved in unexpected directions. Because all the probabilities are set at the individual node level, the behavior of the entire network is difficult to anticipate. However, by repeatedly applying data, and evaluating the network's behavior, probabilities can be tuned so that the system behaves in a manner consistent with human judgments of performance. As on-line data is obtained, the probabilities can be fine-tuned. Moreover, because a student's current status leads to predictions of the classes of actions in given situations, systematic discrepancies can suggest revision of the structure of the network itself.

HYDRIVE employs the same probability-based reasoning that underlies test theory, in an assessment model meant to support instruction as well as evaluation, from the perspective of cognitive psychology, in the context of interactive learning. An articulated cognitive model for performance in the domain provides a coherent framework to move from detailed analysis of discrete actions, to inferences about more general student characteristics. This individual level of information seems necessary to direct instruction to issues that transcend particular problem states and to support broader claims about competence. Since assessment is fundamentally a process of making generalized inferences based on specific information, an approach with test-theoretic roots can contribute to the development of assessment in the ITS world.

Conclusion

Educational test theory has begun to follow multiple paths of progress, to support inference and decision-making from the perspective of contemporary cognitive psychology. Specialists 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. 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. There are many directions to move beyond the simple psychological models and data types of familiar test theory, each presenting its own challenges. Test theorists can play a vital role in this endeavor—not solely as experts at solving inferential problems cast in trait and behavioral terms, but as experts in evidence and inference in school learning problems as cast in psychological frameworks that suit those problems.

References

American Council on the Training of Foreign La nguages. (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.

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.

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.

Brown, J. S., Burton, R. R. & Bell, A. G. (1974). SOPHIE: A sophisticated instructional environment for teaching electronic troubleshooting. *BBN REPORT 2790.* Cambridge, MA: Bolt Beranek and Newman, Inc.

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

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

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. (1982). *Out of the crisis.* Cambridge, MA: Center for Advanced Engineering Study, Massachusetts Institute of Technology.

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.

Embretson, S.E. (Ed.) (1985). *Test design: Developments in psychology and psychometrics.* Orlando: Academic Press.

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., & White, B.Y. (1988). Implicit testing within an intelligent tutoring system. *Machine-Mediated Learning, 2,* 351-372.

Gitomer, D.H., Steinberg, L.S., & Mislevy, R.J. (1995). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 73-101). Hillsdale, NJ: Erlbaum.

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.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J.C. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement: The Buros-Nebraska Symposium on measurement and testing* (Vol. 3) (pp. 41-85). Hillsdale, NJ: Erlbaum.

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

Greeno, J.C. (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., & 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: American Council on Education/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.

Kieras, D.E. (1988). what mental model should be taught: choosing instructional content for complex engineered systems. In M.J. Psotka, L.D. Massey, & S.A. Mutter (Eds.), *Intelligent tutoring systems: Lessons learned* (pp. 85-111). Hillsdale, NJ: Lawrence Erlbaum.

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, D.C.: American Council on Education.

Kuhn, T.S. (1970). *The structure of scientific revolutions* (2nd edition). 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., Eggan, G., Katz, S., & Rao, G. (1992). Possibilities for assessment using computer-based apprenticeship environments. In J. W. Regian and V.J. Shute (Eds.), *Cognitive approaches to automated instruction* (pp. 49-80). Hillsdale, NJ: Lawrence Erlbaum.

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

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

Lewis, C. (1986). Test theory and *Psychometrika*: The past twenty-five years. *Psychometrika, 51*, 11-22.

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.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Martin, J.D., & VanLehn, K. (1993). OLEA: Progress toward a multi-activity, Bayesian student modeler. In S.P. Brna, S. Ohlsson, & H. Pain (Eds.), *Artificial intelligence in education: Proceedings of AI-ED 93* (pp. 410-417). Charlottesville, VA: Association for the Advancement of Computing in Education.

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

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: American Council on Education/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. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R.J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

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. (1988). Toward an explanatory model of document literacy. *Discourse Processes and Press.*

Myford, C.M., & Mislevy, R.J. (1995). *Monitoring and improving a portfolio assessment system.* Center for Performance Assessment Research Report. Princeton, NJ: Educational Testing Service.

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: Kaufmann.

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 Publishers.

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.

Schum, D.A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

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.

Smith, R. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46*, 359- 372.

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: American Council on Education/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.

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

Sternberg, R.J. (1977). *Intelligence, information processing, and analogical reasoning.* New York: Halsted Press.

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. (1989). Problem-solving and cognitive skill acquisition. In M. Posner (Ed.), *The foundations of cognitive science* (pp. 527-580). Cambridge, MA: MIT Press.

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.

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.

Wenger, E. (1987). *Artificial intelligence and tutoring systems.* Los Altos, CA: Morgan Kaufman.

Table 1

Excepts 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	Has sufficient control of the writing system to interpret written language in areas of practical need. ... At times, but not on a consistent basis, the novice-high reader may be able to derive meaning from material at a slightly higher level where context and/or extralinguistic background knowledge are supportive.
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	:

* Based on the *ACTFL proficiency guidelines,* American Council on the Training of Foreign Languages (1989).

Table 2

Skill Requirements for Fractions Items

Item #	Text	1	If Method A used				If Method B used			
			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	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

Table 3

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

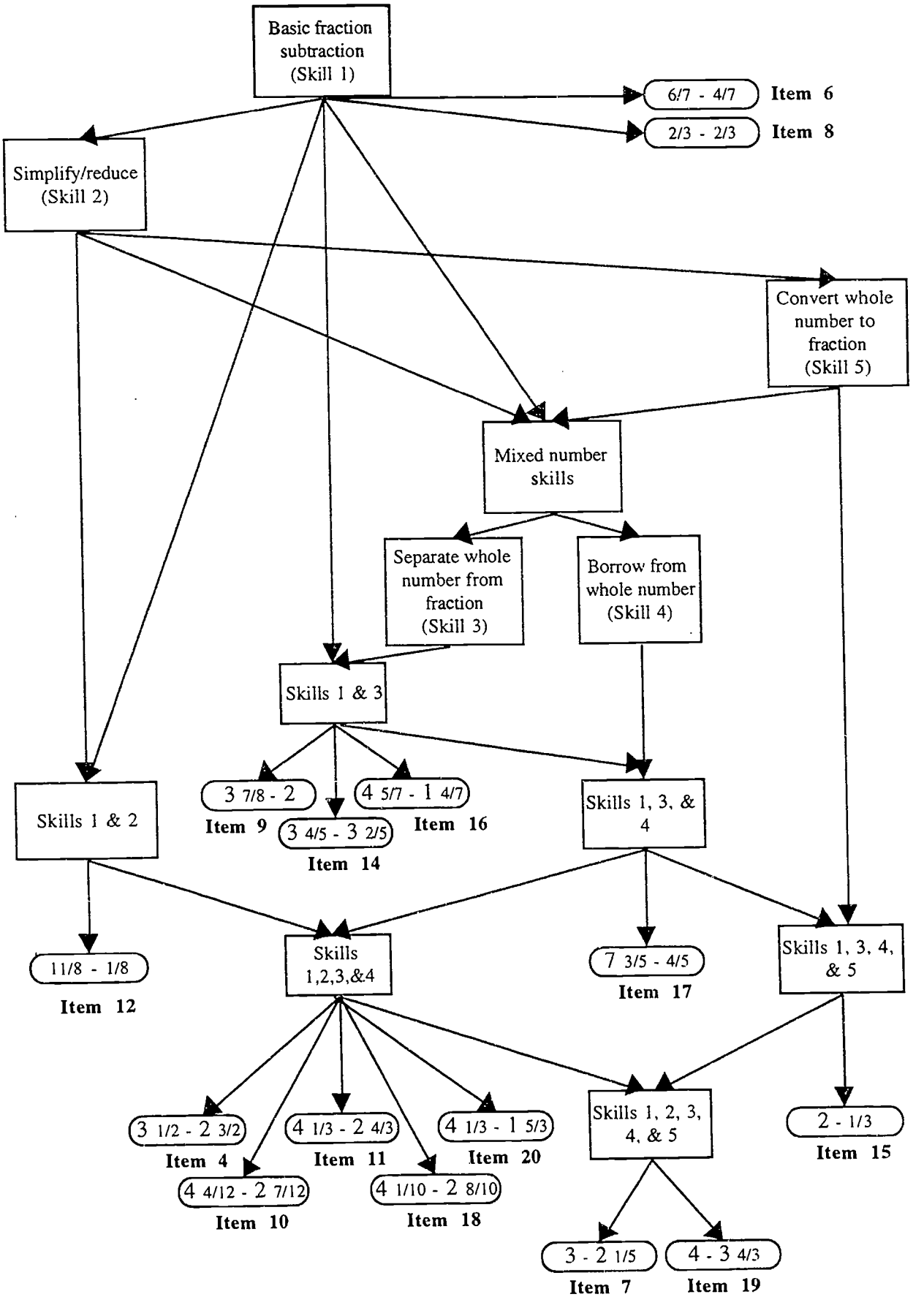
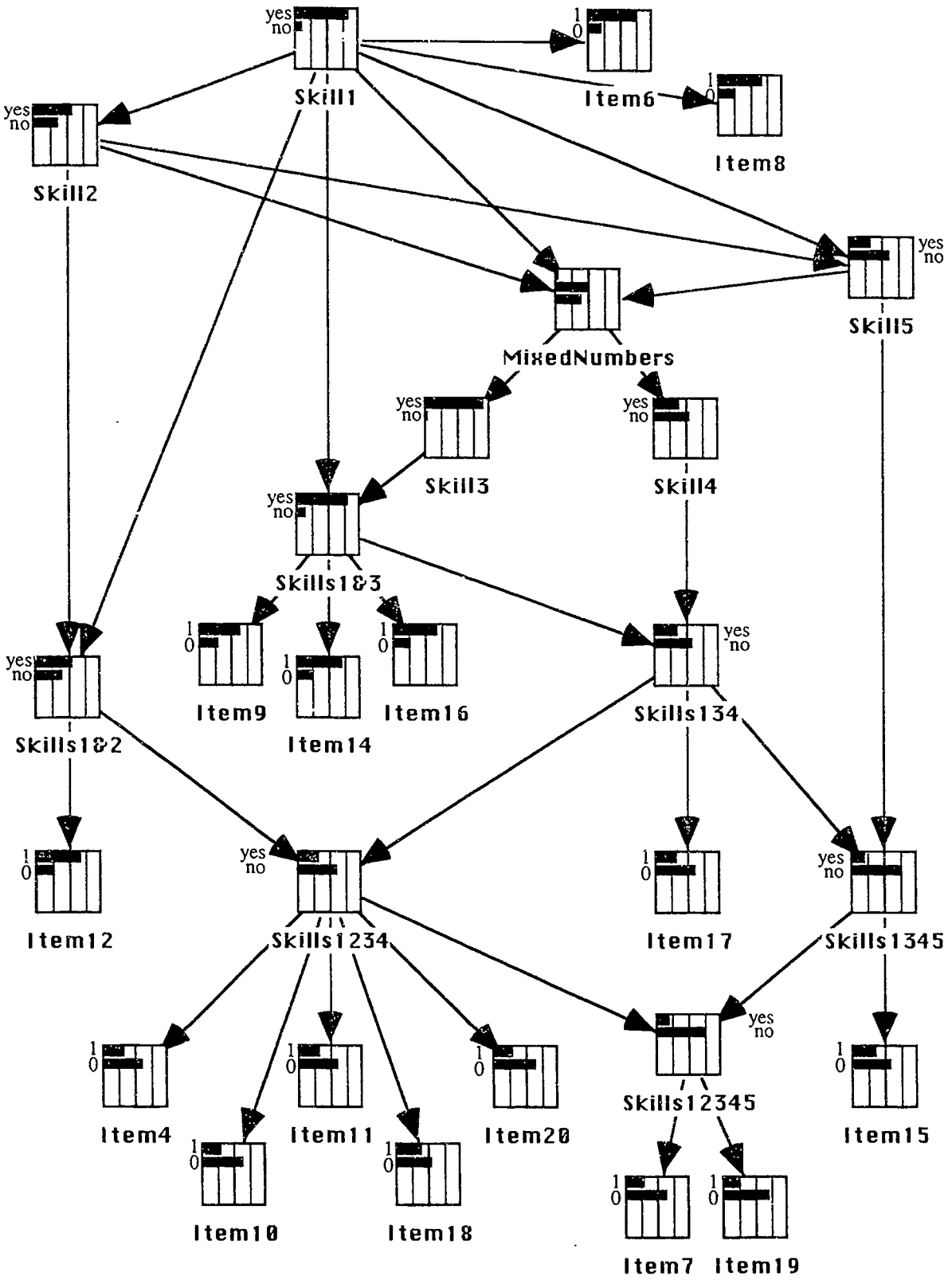


Figure 1

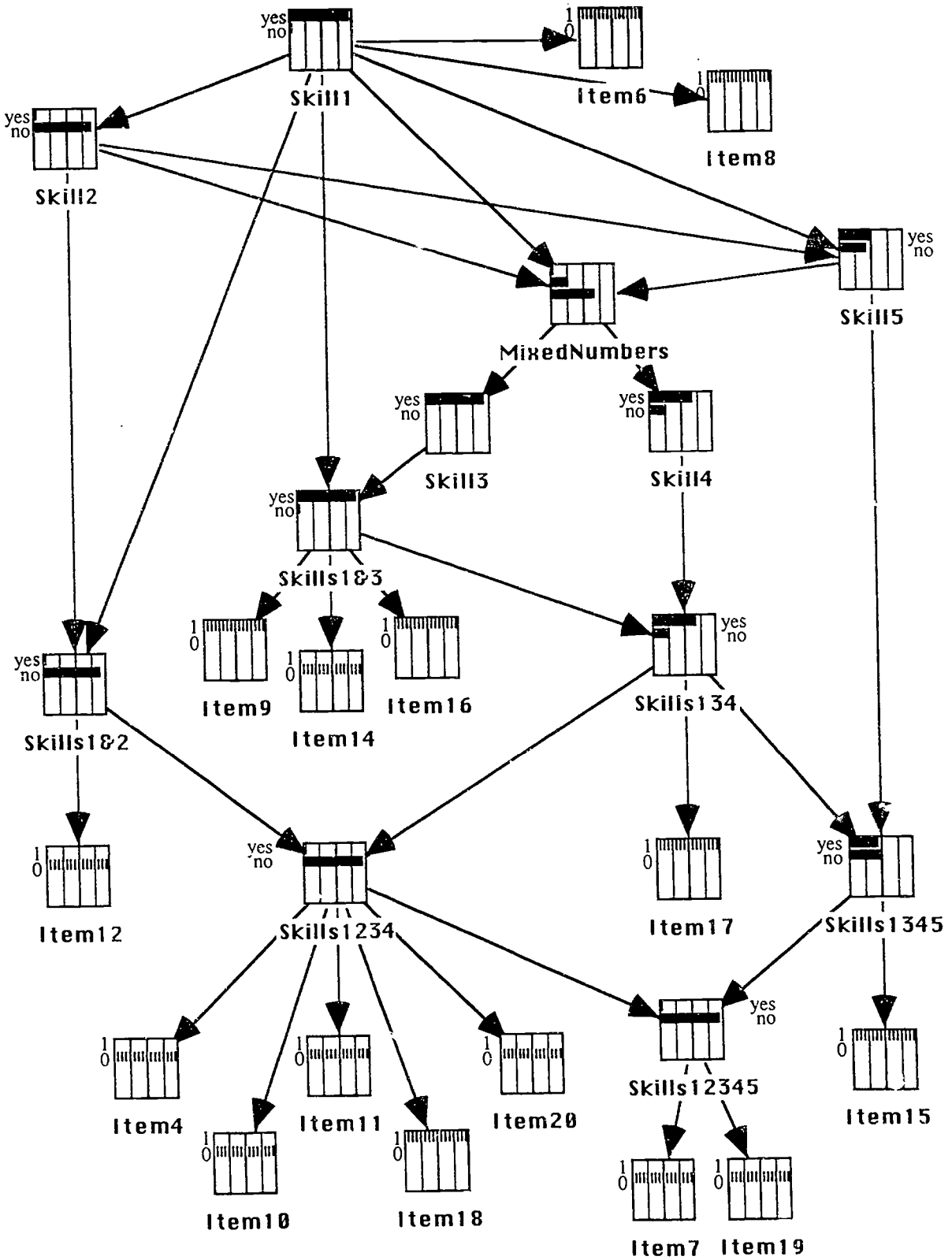
Structure of Inference Network for Method B



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

Figure 2

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 3

Posterior Probabilities for Method B Following Item Responses

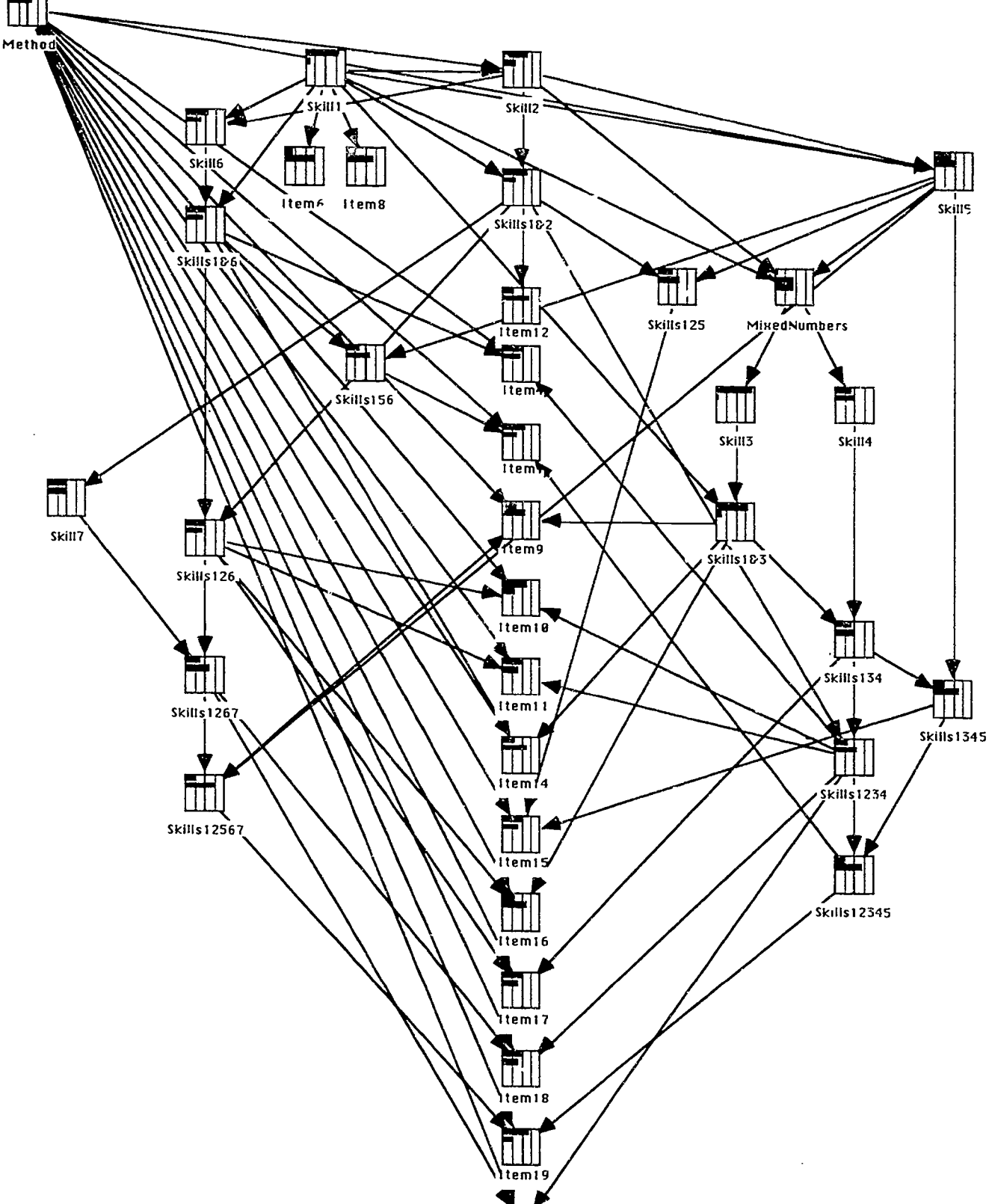
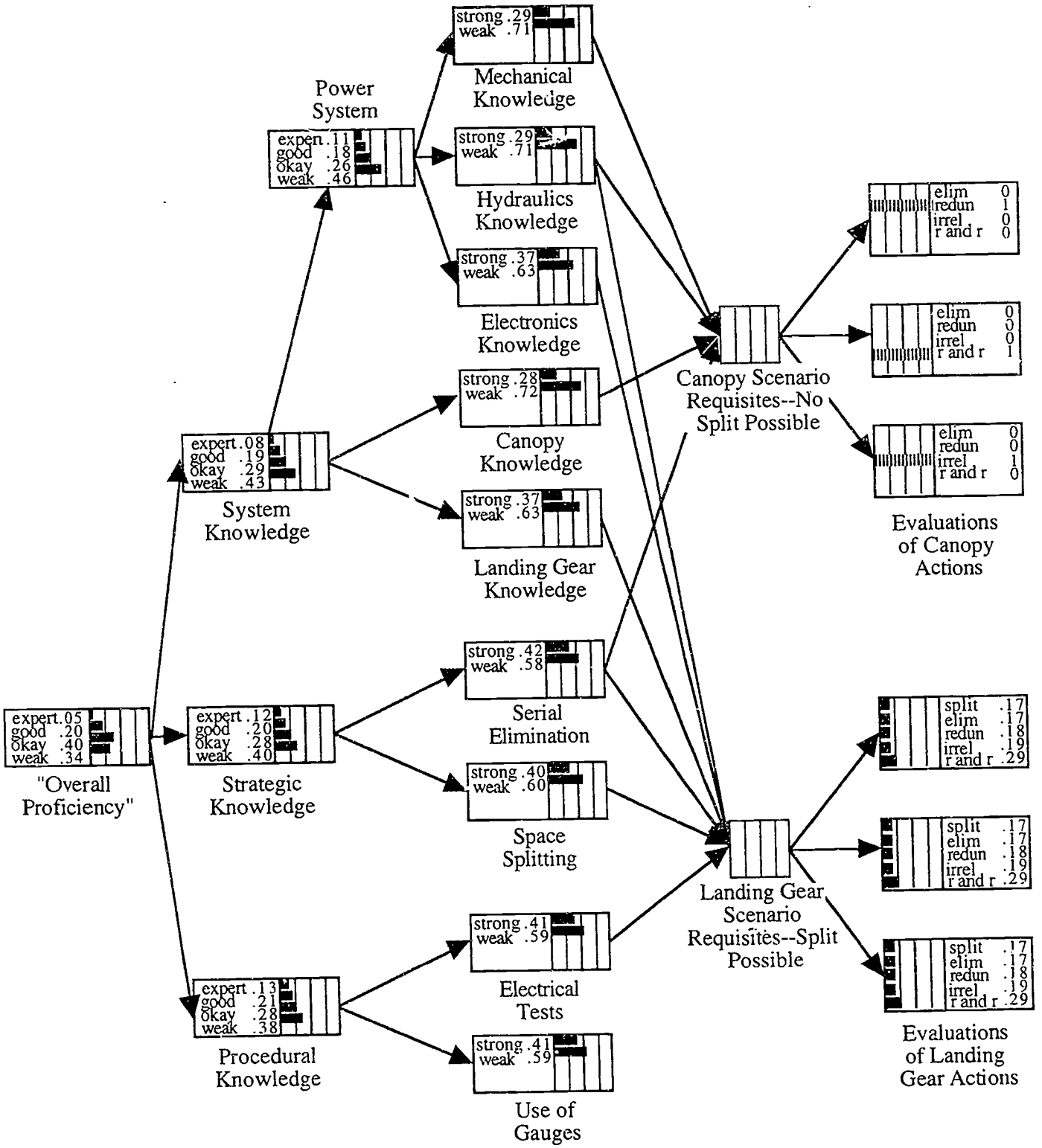


Figure 4

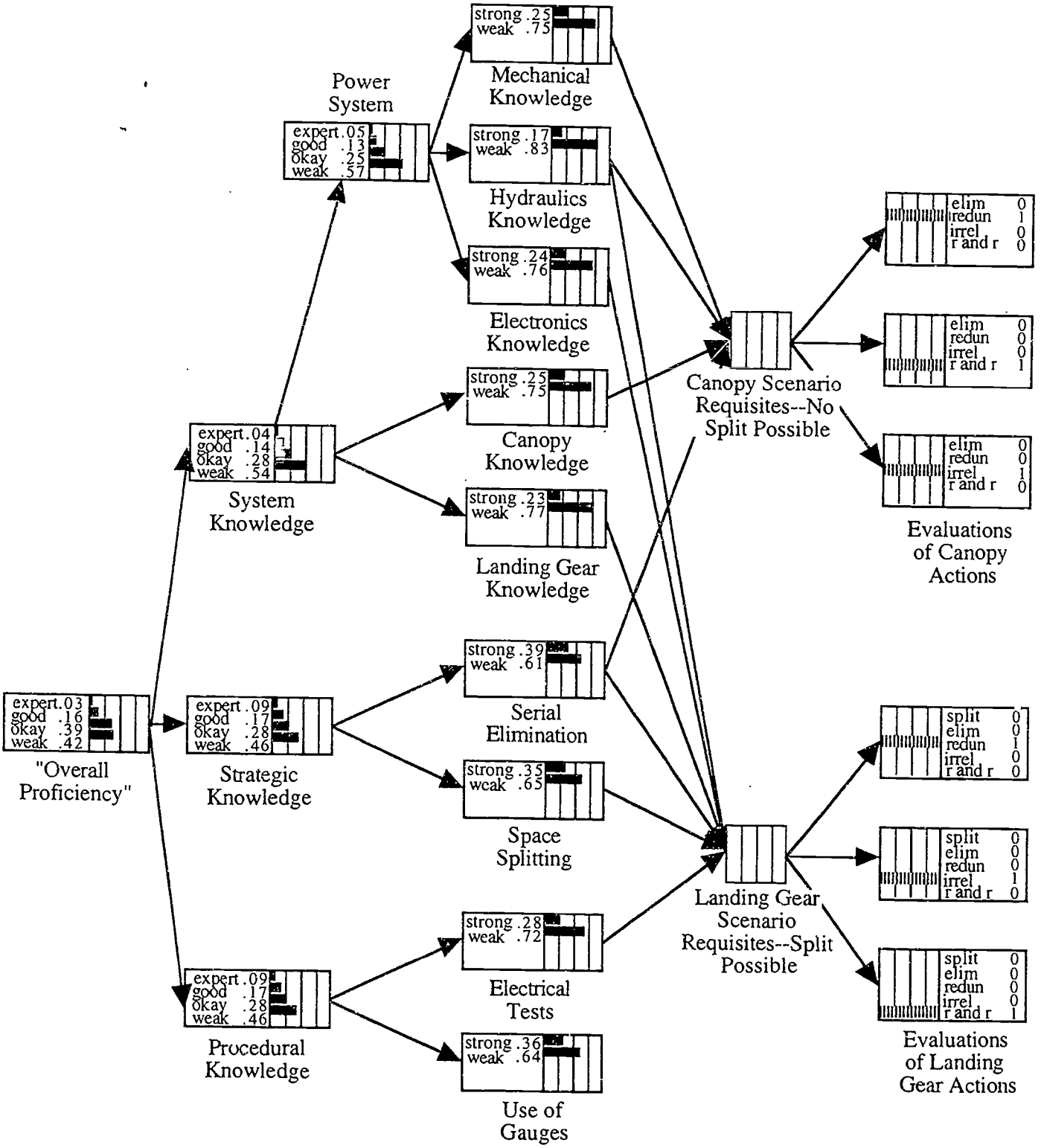
Prior Probabilities in Inference Network for Both Methods Combined



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 5

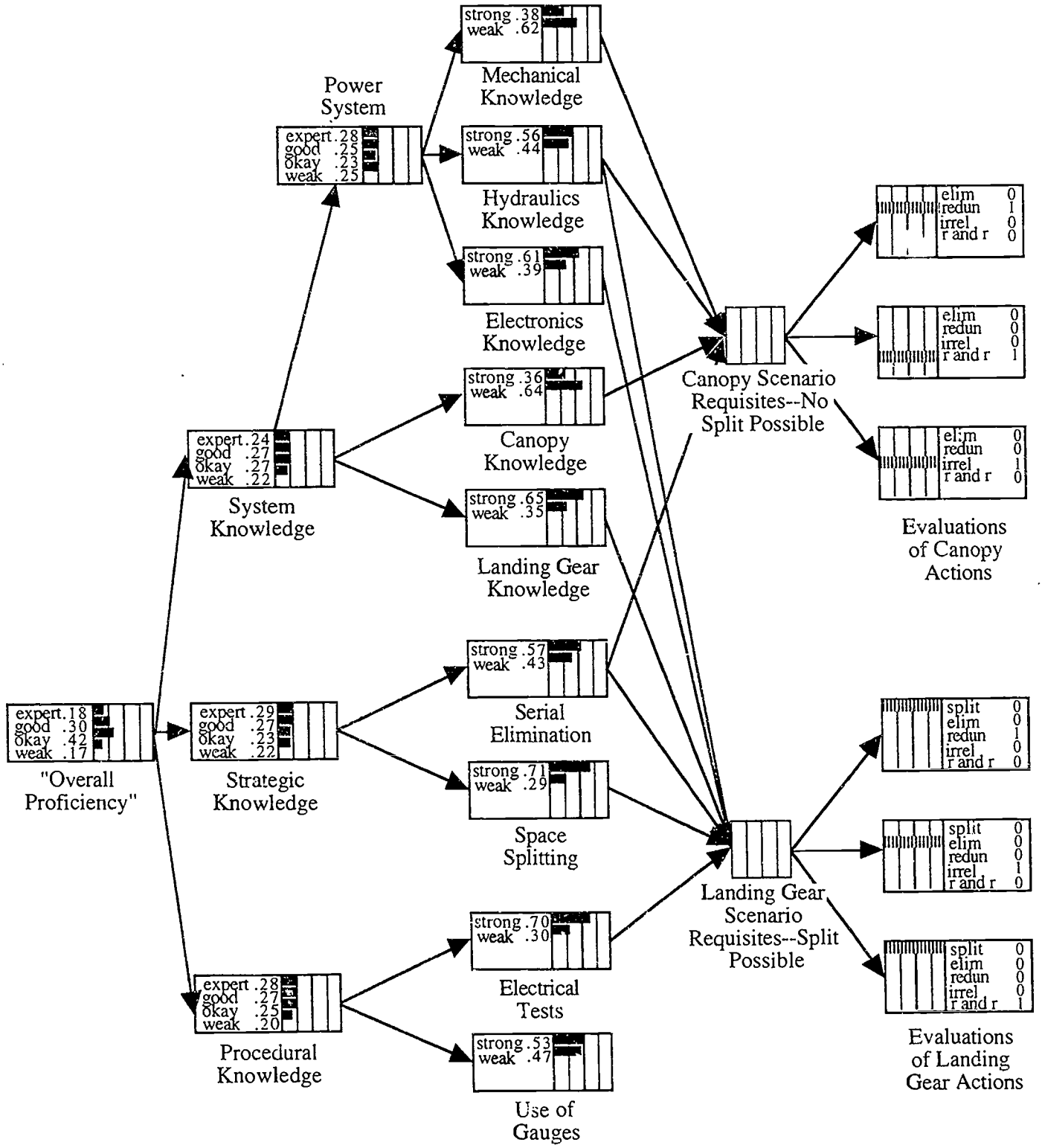
Status of Student Model after Observing Three Inexpert Actions in Canopy Situations



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 6

Status of Student Model after Observing Three Inexpert Actions in Canopy Situations and Three Inexpert Actions in Landing Gear Situations



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 7

Status of Student Model after Observing Three Inexpert Actions in Canopy Situations and Three Expert Actions in Landing Gear Situations

Dr Terry Ackerman
Educational Psychology
260C Education Bldg
University of Illinois
Champaign IL 61801

Dr Robert L Albright
Educational Testing Service
16-C Rosedale Road
Princeton NJ 08541

Dr Terry Allard
Code 3422
Office of Naval Research
800 N Quincy St
Arlington VA 22217-5660

Dr Nancy Allen
Educational Testing Service
Mail Stop 02-T
Princeton NJ 08541

Dr Phipps Arabie
Graduate School of Management
Rutgers University
92 New Street
Newark NJ 07102-1895

Dr Isaac I Bejar
Educational Testing Service
Mail Stop 11-R
Princeton NJ 08541

Dr William O Berry
Director
Life and Environmental Sciences
AFOSR/NL N1
Bldg 410
Bolling AFB DC 20332-6448

Dr Thomas G Bever
Department of Psychology
University of Rochester
River Station
Rochester NY 14627

Dr Menucha Birenbaum
School of Education
Tel Aviv University
Ramat-Aviv 69978 ISRAEL

Dr Bruce Bloxom
Defense Manpower Data Center
99 Pacific St
Suite 155A
Monterey CA 93943-3231

Dr Gwyneth Boodoo
Educational Testing Service
Mail Stop 03-T
Princeton NJ 08541

Dr Richard L Branch
HQ USMEPCOM/MEPCT
2500 Green Bay Road
North Chicago IL 60064

Dr Robert Brennan
American College Testing
2201 North Dodge Street
PO Box 168
Iowa City IA 52243

Dr David V Budescu
Department of Psychology
University of Haifa
Mount Carmel Haifa 31999
ISRAEL

Dr Gregory Candell
CTB/MacMillan/McGraw-Hill
2500 Garden Road
Monterey CA 93940

Dr Susan Chipman
Cognitive Science Program
Office of Naval Research
800 North Quincy Street
Code 3422
Arlington VA 22217-5660

Dr Raymond E Christal
UES LAMP Science Advisor
AL/HRMIL
Brooks AFB TX 78235

Dr Norman Cliff
Department of Psychology
University of Southern California
Los Angeles CA 90089-1061

Dr Nancy S Cole
Educational Testing Service
14-C Rosedale Road
Princeton NJ 08541

Commanding Officer
Naval Research Laboratory
Code 4827
Washington DC 20375-5000

Dr John M Cornwell
Department of Psychology
I/O Psychology Program
Tulane University
New Orleans LA 70118

Dr Linda Curran
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd
Rosslyn VA 22209

Professor Clément Dassa
Faculté des sciences de l'éducation
Département d'études en éducation
et d'administration de l'éducation
CP 6128 succursale A
Montéal Québec
CANADA H3C 3J7

Dr Timothy Davey
American College Testing
2201 North Dodge Street
PO Box 168
Iowa City IA 52243

Dr Charles E Davis
Educational Testing Service
Mail Stop 16-T
Princeton NJ 08541

Dr Ralph J DeAyala
Meas Stat and Eval
Benjamin Bldg Room 1230F
University of Maryland
College Park MD 20742

Director
Life Sciences
Code 3420
Office of Naval Research
Arlington VA 22217-5660

Hei-Ki Dong
BELLCORE
6 Corporate Place
RM: PYA-1K207
PO Box 1320
Piscataway NJ 08855-1320

Dr Neil Dorans
Educational Testing Service
Mail Stop 07-E
Princeton NJ 08541

Dr Fritz Drasgow
University of Illinois
Department of Psychology
603 E Daniel Street
Champaign IL 61820

Defense Tech Information Center
Cameron Station Bldg 5
Alexandria VA 22314
(2 Copies)

Dr Richard Duran
Graduate School of Education
University of California
Santa Barbara CA 93106

Dr Susan Embretson
University of Kansas
Psychology Department
426 Fraser
Lawrence KS 66045

Dr George Engelhard Jr
Division of Educational Studies
Emory University
210 Fishburne Bldg
Atlanta GA 30322

Dr Marshall J Farr
Farr-Sight Co
2520 North Vernon Street
Arlington VA 22207

Dr Leonard Feldt
Lindquist Center for Measurement
University of Iowa
Iowa City IA 52242

Dr Richard L Ferguson
American College Testing
2201 North Dodge Street
PO Box 168
Iowa City IA 52243

Dr Gerhard Fischer
Liebiggasse 5
A 1010 Vienna
AUSTRIA

Dr Myron Fischl
US Army Headquarters
DAPE-HR
The Pentagon
Washington DC 20310-0300

Mr Paul Foley
Navy Personnel R&D Center
San Diego CA 92152-6800

Chair
Department of Computer Science
George Mason University
Fairfax VA 22030

Dr Robert D Gibbons
University of Illinois at Chicago
NPI 909A M/C 913
912 South Wood Street
Chicago IL 60612

Dr Janice Gifford
University of Massachusetts
School of Education
Amherst MA 01003

Dr Robert Glaser
Learning Res & Dev Cntr
University of Pittsburgh
3939 O'Hara Street
Pittsburgh PA 15260

Dr Susan R Goldman
Peabody College
Box 45
Vanderbilt University
Nashville TN 37203

Dr Timothy Goldsmith
Department of Psychology
University of New Mexico
Albuquerque NM 87131

Dr Sherrie Gott
AFHRL/MOMJ
Brooks AFB TX 78235-5601

Dr Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore MD 21218

Professor Edward Haertel
School of Education
Stanford University
Stanford CA 94305-3096

Dr Ronald K Hambleton
University of Massachusetts
Lab of Psychom & Eval Res
Hills South Room 152
Amherst MA 01003

Dr Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign IL 61820

Dr Patrick R Harrison
Computer Science Department
US Naval Academy
Annapolis MD 21402-5002

Ms Rebecca Hetter
Navy Personnel R&D Center
Code 13
San Diego CA 92152-6800

Dr Thomas M Hirsch
American College Testing
2201 North Dodge Street
PO Box 168
Iowa City IA 52243

Professor Paul W Holland
Div of Educ Psych &
Quant Methods Prog
Graduate School of Education
4511 Tolman Hall
University of California-Berkeley
Berkeley CA 94720

Professor Lutz F Hornke
Institut fur Psychologie
RWTH Aachen
Jaegerstrasse 17/19
D-5100 Aachen
WEST GERMANY

Ms Julia S Hough
Cambridge University Press
40 West 20th Street
New York NY 10011

Dr William Howell
Chief Scientist
AFHRL/CA
Brooks AFB TX 78235-5601

Dr Huynh Huynh
College of Education
University of South Carolina
Columbia SC 29208

Dr Martin J Ippel
Center for the Study of
Education & Instruction
Leiden University
PO Box 9555
2300 RB Leiden
THE NETHERLANDS

Dr Robert Jannarone
Elec.and Computer Eng Dept
University of South Carolina
Columbia SC 29208

Dr Kumar Joag-dev
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign IL 61820

Professor Douglas H Jones
Grad Sch of Management
Rutgers The State University
NJ Newark NJ 07102

Dr Brian Junker
Carnegie-Mellon University
Department of Statistics
Pittsburgh PA 15213

Dr Marcel Just
Carnegie-Mellon University
Department of Psychology
Schenley Park
Pittsburgh PA 15213

Dr J L Kaiwi
Code 442/JK
Naval Ocean Systems Center
San Diego CA 92152-5000

Dr Michael Kaplan
Office of Basic Research
US Army Research Institute
5001 Eisenhower Avenue
Alexandria VA 22333-5600

Dr Jeremy Kilpatrick
Dept of Mathematics Education
105 Aderhold Hall
University of Georgia
Athens GA 30602

Ms Hae-Rim Kim
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign IL 61820

Dr. Jwa-keun Kim
Department of Psychology
Middle Tennessee State University
Murfreesboro TN 37132

Dr Sung-Hoon Kim
KEDI
92-6 Umyeon-Dong
Seocho-Gu
Seoul
SOUTH KOREA

Dr G Gage Kingsbury
Portland Public Schools
Res & Eval Department
501 North Dixon Street
PO Box 3107
Portland OR 97209-3107

Dr William Koch
Box 7246
Meas & Eval Center
University of Texas-Austin
Austin TX 78703

Dr James Kraatz
Computer-based Ed Res Lab
University of Illinois
Urbana IL 61801

Dr Patrick Kyllonen
AFHRL/MOEL
Brooks AFB TX 78235

Ms Carolyn Laney
1515 Spencerville Rod
Spencerville MD 20868

Richard Lanterman
Commandant (G-PWP)
US Coast Guard
2100 Second Street SW
Washington DC 20593-0001

Dr Michael Levine
Educational Psychology
210 Education Building
1310 South Sixth Street
Univ of IL at Urbana-Champaign
Champaign IL 61820-6990

Dr Charles Lewis
Educational Testing Service
Mail Stop 03-T
Princeton NJ 08541-0001

Mr Hsin-hung Li
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign IL 61820

Library
Naval Training Systems Center
12350 Research Parkway
Orlando FL 32826-3224

Dr Marcia C Linn
Graduate School of Education
EMST
Tolman Hall
University of California
Berkeley CA 94720

Dr Robert L Linn
Campus Box 249
University of Colorado
Boulder CO 80309-0249

Dr Richard Luecht
American College Testing
2201 North Dodge Street
PO Box 168
Iowa City IA 52243

Dr. George Macready
Professor
Dept of Meas., Stat. & Eval.
College of Education
Room 1230C. Benjamin Bldg.
University of Maryland
College Park MD 20742

Dr Evans Mandes
George Mason University
4400 University Drive
Fairfax VA 22030

Dr Paul Mayberry
Center for Naval Analysis
4401 Ford Avenue
PO Box 16268
Alexandria VA 22302-0268

Dr James R McBride
HumRRO
6430 Elmhurst Drive
San Diego CA 92120

Mr Christopher McCusker
University of Illinois
Department of Psychology
603 E Daniel Street
Champaign IL 61820

Dr Joseph McLachlan
Navy Pers Res & Dev Cntr
Code 14
San Diego CA 92152-6800

Alan Mead
c/o Dr Michael Levine
Educational Psychology
210 Education Bldg
University of Illinois
Champaign IL 61801

Dr Timothy Miller
American College Testing
2201 North Dodge Street
PO Box 168
Iowa City IA 52243

Dr Robert Mislevy
Educational Testing Service
Mail Stop 03-T
Princeton NJ 08541

Dr Ivo Molenar
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Grote Kruisstraat 2/1
9712 TS Groningen
The NETHERLANDS

Dr Eiji Muraki
Educational Testing Service
Mail Stop 02-T
Princeton NJ 08541

Dr Ratna Nandakumar
Educational Studies
Willard Hall Room 213E
University of Delaware
Newark DE 19716

Acad Prog & Research Branch
Naval Tech Training Command
Code N-62
NAS Memphis (75)
Millington TN 30854

Dr W Alan Nicewander
American College Testing
2201 North Dodge Street
PO Box 168
Iowa City IA 52243

Head
Personnel Systems Department
NPRDC (Code 12)
San Diego CA 92152-6800

Director
Training Systems Department
NPRDC (Code 14)
San Diego CA 92152-6800

Library NPRDC
Code 041
San Diego CA 92152-6800

Librarian
Naval Cntr for Applied Research
in Artificial Intelligence
Naval Research Laboratory
Code 5510
Washington DC 20375-5000

Office of Naval Research
Code 3422
800 N Quincy Street
Arlington VA 22217-5660

ONR Resident Representative
New York City
33 Third Avenue - Lower Le
New York NY 10003-9998

Special Asst for Res Management
Chief of Naval Personnel
(PERS-O1JT)
Department of the Navy
Washington DC 20350-2000

Dr Judith Orasanu
NASA Ames Research Center
Mail Stop 239-1
Moffett Field CA 94035

Dr Peter J Pashley
Law School Admission Services
PO Box 40
Newtown PA 18940-0040

Wayne M Patience
American Council on Education
GED Testing Service Suite 20
One Dupont Circle NW
Washington DC 20036

Dept of Administrative Sciences
Code 54
Naval Postgraduate School
Monterey CA 93943-5026

Dr Peter Pirolli
School of Education
University of California
Berkeley CA 94720

Dr Mark D Reckase
American College Testing
2201 North Dodge Street
PO Box 168
Iowa City IA 52243

Mr Steve Reise
Department of Psychology
University of California
Riverside CA 92521

Mr Louis Roussos
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign IL 61820

Dr Donald Rubin
Statistics Department
Science Center Room 608
1 Oxford Street
Harvard University
Cambridge MA 02138

Dr Fumiko Samejima
Department of Psychology
University of Tennessee
310B Austin Peay Bldg
Knoxville TN 37966-0900

Dr Mary Schratz
4100 Parkside
Carlsbad CA 92008

Mr Robert Semmes
N218 Elliott Hall
Department of Psychology
University of Minnesota
Minneapolis MN 55455-0344

Dr Valerie L Shalin
Dept of Industrial Engineering
State University of New York
342 Lawrence D Bell Hall
Buffalo NY 14260

Mr Richard J Shavelson
Graduate School of Education
University of California
Santa Barbara CA 93106

Kathleen Sheehan
Educational Testing Service
Mail Stop 03-T
Princeton NJ 08541

Dr Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujisawa 251
JAPAN

Dr Randall Shumaker
Naval Research Laboratory
Code 5500
4555 Overlook Avenue SW
Washington DC 20375-5000

Dr Judy Spray
American College Testing
2201 North Dodge Street
PO Box 168
Iowa City IA 52243

Dr Martha Stocking
Educational Testing Service
Mail Stop 03-T
Princeton NJ 08541

Dr William Stout
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St
Champaign IL 61820

Dr Kikumi Tatsuoka
Educational Testing Service
Mail Stop 03-T
Princeton NJ 08541

Dr David Thissen
Psychometric Laboratory
CB# 3270 Davie Hall
University of North Carolina
Chapel Hill NC 27599-3270

Mr Thomas J Thomas
Federal Express Corporation
Human Resource Development
3035 Director Row Suite 501
Memphis TN 38131

Mr Gary Thomasson
University of Illinois
Educational Psychology
Champaign IL 61820

Dr Howard Wainer
Educational Testing Service
15-T Rosedale Road
Princeton NJ 08541

Elizabeth Wald
Office of Naval Technology
Code 227
800 North Quincy Street
Arlington VA 22217-5000

Dr Michael T Waller
Univ of Wisconsin-Milwaukee
Educ Psychology Department
Box 413
Milwaukee WI 53201

Dr Ming-Mei Wang
Educational Testing Service
Mail Stop 03-T
Princeton NJ 08541

Dr Thomas A Warm
FAA Academy
PO Box 25082
Oklahoma City OK 73125

Dr David J Weiss
N660 Elliott Hall
University of Minnesota
75 E River Road
Minneapolis MN 55455-0344

Dr Douglas Wetzel
Code 15
Navy Personnel R&D Center
San Diego CA 92152-6800

German Military Representative
Personalstammamt
Koelner Str 262
D-5000 Koeln 90
WEST GERMANY

Dr David Wiley
Sch of Educ and Social Policy
Northwestern University
Evanston IL 60208

Dr Bruce Williams
Dept of Educational Psychology
University of Illinois
Urbana IL 61801

Dr Mark Wilson
School of Education
University of California
Berkeley CA 94720

Dr Eugene Winograd
Department of Psychology
Emory University
Atlanta GA 30322

Martin F Wiskoff
PERSEREC
99 Pacific Street
Suite 4556
Monterey CA 93940

Mr John H Wolfe
Navy Personnel R&D Center
San Diego CA 92152-6800

Dr Kentaro Yamamoto
Educational Testing Service
Mail Stop 03-T
Princeton NJ 08541

Duanli Yan
Educational Testing Service
Mail Stop 03-T
Princeton NJ 08541

Dr Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey CA 93940

Dr Joseph L Young
National Science Foundation
Room 320
1800 G Street NW
Washington DC 20550