Evidence and Inference in Educational Assessment

CSE Technical Report 414

Robert J. Mislevy
Educational Testing Service

May 1996

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

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

The views expressed here are those of the author and do not imply any official endorsement by any organizations funding this work.

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

EVIDENCE AND INFERENCE IN EDUCATIONAL ASSESSMENT

Robert J. Mislevy
Educational Testing Service

Abstract

Educational assessment concerns inference about students' knowledge, skills, and accomplishments. Because data are never so comprehensive and unequivocal as to ensure certitude, test theory evolved in part to address questions of weight, coverage, and import of data. The resulting concepts and techniques can be viewed as applications of more general principles for inference in the presence of uncertainty. Issues of evidence and inference in educational assessment are discussed from this perspective.

Key words: Bayesian inference networks, cognitive psychology, evidence, inference, performance assessment, probability, psychometrics, test theory

EVIDENCE AND INFERENCE IN EDUCATIONAL ASSESSMENT[1]

Robert J. Mislevy
Educational Testing Service

Probability isn't really about numbers; it's about the structure of reasoning.
Glenn Shafer (quoted in Pearl, 1988)

Introduction

Harold Gulliksen, reviewing the field of *Measurement of Learning and Mental Abilities* at the 25th anniversary of the Psychometric Society in 1961, described "the central problem of test theory" as "the relation between the ability of the individual and his [or her] observed score on the test" (Gulliksen, 1961). Twenty-five years later, at the 50th anniversary, Charles Lewis observed that "much of the recent progress in test theory has been made by treating the study of the relationship between responses to a set of test items and a hypothesized trait (or traits) of an individual as a problem of statistical inference" (Lewis, 1986). This trend represents practical progress to be sure, providing solutions to formerly intractable problems such as tailoring tests to individual examinees (e.g., Lord, 1980, chapter 10) and sorting out relationships in patterns of achievement in hierarchical schooling systems (e.g., Aitkin & Longford, 1986).

Perhaps more importantly in the long run, it represents a certain progress in understanding. The early literature on test theory blurred the distinction between models for students' knowledge or accomplishments on the one hand, and, on the other, an observer's state of knowledge about the forms and parameters of these models. The statistical developments Lewis spoke of helped researchers explicate the evidence that test data convey for assessment problems framed under trait and behaviorist psychological conceptions of abilities. Ironically, the very success

[1] Presidential address to the Psychometric Society, presented June 25, 1994, in Champaign, Illinois. I am grateful for comments and suggestions from Henry Braun, Drew Gitomer, Richard Patz, Jonathan Troper, and Howard Wainer.

of statistical reasoning for assessment problems cast under the trait and behaviorist paradigms gave rise to a misconception that statistical reasoning applies to assessment framed *only* within those paradigms.

We can, however, view test theory as the application of principles that have evolved over hundreds of years in many fields, to deal with such pervasive problems as multistage inference and multiple sources of disparate evidence. While recent developments in cognitive and educational psychology may suggest student models and observational strategies quite different from those employed by, say, Spearman, Thurstone, and Thorndike, practical work under alternative perspectives inevitably faces these same general problems in some form. The same general principles of inference—central among them the concepts and tools of mathematical probability—can help explicate relationships between evidence and inference for a broader discourse about students' knowledge, learning, and accomplishments than is traditionally associated with standard test theory and standardized achievement tests. This paper aims to elaborate this claim and to illustrate points with vignettes from current projects.

The following section reviews basic ideas about evidence and inference, drawing in part from Schum's (1987) monograph, *Evidence and Inference for the Intelligence Analyst*. Jurist John Henry Wigmore's contributions to understanding the structure of complex bodies of evidence and evidentiary arguments are then discussed (Anderson & Twining, 1991; Wigmore, 1937) with reference to analogous problems in jurisprudence and assessment. Conceptual machinery from mathematical probability-based reasoning that can be applied to these structures is then considered. A series of examples uses this approach to structure inference concerning proportional reasoning, mixed-number subtraction, foreign-language learning, and accomplishment in a studio art program. The focus in each case is modeling evidentiary reasoning, through an inferential model built around a psychological model for competence in the domain. The interplay between probability-based reasoning *within* a model and nonmathematical reasoning *about* the model is then discussed; the former provides a framework for reasoning through the complexities Wigmore described, the latter emphasizes a perspective of criticizing and improving that framework.

Evidence and Inference

> Questions of evidence are continually presenting themselves to every human being, every day, and almost every waking hour, of his life. . . . Whether the leg of mutton now on the spit be roasted enough, is question of evidence . . . which the cook decides upon in the cook way, as if by instinct; deciding upon evidence, as Monsieur Jourdan talked prose, without having ever heard of any such word, perhaps, in the whole course of her life.
>
> Jeremy Bentham, 1827, pp. 18-19

Data versus Evidence

Inference is reasoning from what we know and what we observe to explanations, conclusions, or predictions. We always reason in the presence of uncertainty. The information we work with is typically incomplete, inconclusive, amenable to more than one explanation. We attempt to establish the weight and coverage of evidence in what we observe. But the very first question we must address is "Evidence about what?" Schum (1987, p. 16) stresses the crucial distinction between *data* and *evidence*: "A datum becomes evidence in some analytic problem when its *relevance* to one or more hypotheses being considered is established. . . . Evidence is relevant on some hypothesis [conjecture] if it either increases or decreases the likeliness of the hypothesis. Without hypotheses, the relevance of no datum could be established." The same data can thus prove conclusive for some inferences, but barely suggestive for others; it can provide complete coverage for some inferences, yet miss core issues of others; it can constitute direct evidence for some inferences and indirect evidence for others, yet be wholly irrelevant to still others.

Conjectures, and the understanding of what constitutes evidence about them, emanate from the variables, concepts, and relationships of the field within which reasoning is taking place—the paradigm, to use Kuhn's (1970) term. Educational assessments provide data such as written essays, correct and incorrect marks on answer sheets, presentations of projects, or students' explanations of their problem solutions. These data become evidence only with respect to conjectures about students and their work—conjectures constructed around notions of the character and acquisition of knowledge and skill, and shaped by the purpose of the assessment and the nature of the inference required. For example:

- From a *behavioral* perspective, the focus is on chances of success in a domain of relevant tasks. A student is characterized in terms of "overall proficiency" in the domain in terms of, say, the score that would be expected if she were administered all tasks in the domain, and conjectures would concern her level of proficiency in relation to the tasks themselves or to other students, or her behavior in other situations. Responses to a sample of tasks constitutes direct evidence for a conjecture about proficiency so construed.

- From an *information processing* perspective, competence is construed in terms of "production rules," and conjectures concern the sets of production rules (production systems) students have at their disposal. A production rule comprises descriptions of conditions which, when recognized, trigger actions. An example is "**smaller-from-larger-when-borrowed-from**: When there are two borrows in a row, the student does the first one correctly, but for the second one she does not borrow; instead she subtracts the smaller from the larger digit—e.g., 824-157=747" (VanLehn, 1990, p. 228). Individual production rules can be correct or erroneous; a given production system might handle certain features of the substantive domain correctly but miss others.

- From a *constructivist* perspective, a student comes to understand the important attributes and relations of specific contexts and circumstances (including social circumstances), and through wider experiences extends, connects, and generalizes the patterns so that they may be applied more broadly and more effectively. Conjectures concern the degree to which a student has developed useful knowledge, both within and across particular contexts and circumstances, and the nature of that knowledge (including, for example, the kinds of meaning the student can construct in new situations).

This presentation does not argue that any of these perspectives represents "the truth." All are constructions, organized around patterns that have been perceived in aspects of human learning and problem solving. Each can be useful in certain circumstances to improve learning and problem solving, much as wave and particle models for atomic phenomena are each advantageous for certain physics problems. Our concern is that practical work under any psychological perspective must proceed with less than perfect knowledge. To this end, examples of evidentiary problems in assessment will be illustrated with examples from all three perspectives.

Kinds of Inference

Schum (1987) distinguishes deductive, inductive, and abductive reasoning, all of which play essential and interlocking roles in educational assessment:

- *Deductive reasoning* flows from generals to particulars, within an established framework of relationships among variables—from causes to effects, from diseases to symptoms, from the way a crime is committed to the evidence likely to be found at the scene, from a student's knowledge and skills to observable behavior. Under a given state of affairs, what are the likely outcomes? Formal logic includes instances of conclusive deductive reasoning; accepting "A implies B" and learning "not B," we conclude "not A" with certainty. In practice, deductive reasoning is often probabilistic; under different states, various possibilities become more or less likely but not completely determined.

- *Inductive reasoning* flows in the opposite direction, also within an established framework of relationships—from effects to possible causes, from symptoms to probable diseases, from students' solutions or patterns of solutions to likely configurations of knowledge and skill. Given outcomes, what state of affairs may have produced them?

- *Abductive reasoning* (a term coined by the philosopher Charles S. Peirce) proceeds from observations to new hypotheses, new variables, or new relationships among variables. "Such a 'bottom-up' process certainly appears similar to induction; but there is an argument that such reasoning is, in fact, different from induction since an existing hypothesis collection is enlarged in the process. Relevant evidentiary tests of this new hypothesis are then *deductively* inferred from the new hypothesis" (Schum, 1987, p. 20; emphasis original).

The theories and explanations of a field suggest the structure through which deductive reasoning flows. Inductive and abductive reasoning depend likewise critically on the same structures, as the task is to speculate on circumstances that, when their consequences are projected deductively, lead plausibly to the evidence at hand. Determining promising possibilities, we reason deductively to other likely consequences—potential sources of corroborating or disconfirming evidence for our conjectures, by means of which we may further develop our understanding (Lakatos, 1970).

A detective at the scene of a crime reasons abductively to reconstruct the essentials and principals of the event. Anything he sees, in light of a career of experience, can suggest possibilities; ways things might have happened that, reasoning deductively, could have produced the present state of affairs (e.g., documents, eyewitness reports, physical evidence). Given tentative hypotheses, does inductive reasoning from other observations conflict or fit in? When they conflict, does their juxtaposition spark a new hypothesis? A successful investigation leads to a plausible explanation of the case, that, reasoning deductively, appears to lead convincingly to the data at hand. This is the "theory of the case" the prosecution brings to trial.

Severely limited in time and place, a jury cannot "begin at the beginning" in the same way the detective did. Their charge is to decide whether the mass of evidence the prosecution presents to support this particular hypothesis is sufficiently credible, or whether it falls short when the defense's rebuttals and alternative explanations are considered. The jury addresses a problem of inductive inference—"Does the evidentiary fact point to the desired conclusion (not as the only rational inference, but) as the inference (or explanation) most plausible or most natural out of the various ones that are conceivable?" (Wigmore, 1937, p. 25)—within a framework constructed only through substantial abductive inference on the part of the investigator and the prosecution. Even though the detective may have more information and better insight than the jury ("I *know* the butler did it, but I just can't prove it yet"), the credibility of the legal system is enhanced by this separation: The decision is made on the basis of public presentation of evidence and argument, by different people from those who gathered the evidence and structured the inferential framework.

Probability-Based Reasoning

> According to the assumption of situated cognition, most cognitive activity occurs in direct interaction with a situation, rather than being mediated by cognitive representations. Cognitive representations play a role when something goes wrong. They are resources that humans have for dealing with situations when their more direct connections with objects and persons are not working well. . . . The capabilities that we characterize as critical thinking, then, need to include recognition of circumstances when reflection and evaluation might be helpful in overcoming some difficulty that has emerged in the normal course of activity or conversation.
>
> Greeno, 1989, p. 139

We do not build probability models for most of the reasoning we do, either in our jobs or our everyday lives. We continually reason deductively, inductively, and abductively, to be sure, but not through explicit formal models. Why not? Partly because we use heuristics, which, though suboptimal (e.g., Kahneman, Slovic, &

Tversky, 1982), generally suffice for our purposes. More importantly, because much of our reasoning concerns domains we know something about. Greeno (1989, p. 139) continues, "rather than assimilation of information, concepts, and procedures, we can consider learning in a domain as becoming able to think with and about the information, concepts, and procedures of the domain. This includes coming to know the generative principles of the domain, that is, learning what makes the information and procedures of the domain work, rather than simply learning what they are." Attending to the right features of a situation and reasoning through the right relationships, informally or even unconsciously, provides some robustness against suboptimal use of available information within that structure.

Some robustness, but not invincibility. Heuristics, habits, rules of thumb, standards of proof, and typical operating procedures guide practice in substantive domains, more or less in response to what seems to have worked in past and what seems to have led to trouble. This inferential machinery co-evolves with, and is intimately intertwined with, the problems, the concepts, the constraints, and the methodologies of the field (Kuhn, 1970, p. 109). Difficulties arise when inferential problems become so complex that the usual heuristics fail, when the costs of unexamined standard practices become exorbitant, or when novel problems appear. It is in these situations that more generally framed and formally developed systems of inference provide their greatest value.

Given key concepts and relationships, inferential objectives, and data, how *should* reasoning proceed? How can we characterize the nature and force of persuasion a mass of data conveys about a target inference? Workers in every field have had to address these questions as they arise with the kinds of inferences and the kinds of evidence they customarily address. Currently, the promise of computerized expert systems has sparked interest in principles of inference at a level that might transcend the particulars of fields and problems. Historically, this quest has received most attention in the fields of statistics (unsurprisingly), philosophy, and jurisprudence. In the sequel we focus on the concepts and the uses of probability-based reasoning.

Two traditions of "probability" have arisen over time: mathematical or Pascalian (after Blaise Pascal) probability, and epistemic or Baconian (after Francis Bacon) probability. Those of us in test theory are more familiar with Pascalian probability. For our purposes, the essential elements are a specified space of outcomes, or sample space; a space of parameters, or variables that determine how likely outcomes are; and a function that specifies the probabilities of "Pascalian events," or subsets of the sample space, given values of parameters. Probabilities are numbers that satisfy the following requirements: (a) an event's probability is greater than or equal to 0, (b) the probability of the event that includes all possible outcomes is 1, and (c) the probability of an event defined as the union of a collection of disjoint events is the sum of their individual probabilities (Kolmogorov, 1950); they correspond to strength of belief. It is portentous that given parameter values, we can express the relative chances of a Pascalian event as compared to any other events; and given an event, we can express the relative plausibility of a given parameter value as compared to any other parameter value. We shall have more to say about this aspect of Pascalian probability-based inference below.

In contrast, a "Baconian event" is closer to the everyday notion of "something that has happened." Baconian probability refers to a conviction of belief or persuasion, without necessary reference to a numerical characterization of its strength, a specifiable sample space (things that "might have happened," in addition to "what did happen"), a parameter space (potential "true states of affairs" that might have led to the observed event), or functions that explicate the relationships between what is observed and what is inferred. We may nevertheless be able to say that given the evidence, we feel that one conjecture is more likely than another (Cohen, 1977). We find ourselves mildly or strongly convinced of a conjecture given a body of data, and we may be able to lay out arguments that persuade us or give us pause. This Baconian perspective underlies much judicial evidentiary reasoning, and from this perspective, John Henry Wigmore, Dean of Evidence at Northwestern University in the first third of the century, was able to identify, if not resolve, some central inferential challenges.

Wigmore on Evidence

Wigmore, like Jeremy Bentham a hundred years before him, was troubled by the agglomeration of "rules of evidence" that had evolved in Anglo-American law over the centuries. Each rule, specifying particular kinds or aspects of information that may or may not be introduced to jurors as evidence in a case, is intended to reduce the chances of some presumed inferential error. Beyond the fact that certain rules offend sensibility (Quakers could not give testimony in some jurisdictions because they refused to swear an oath of truthfulness), Wigmore felt that what was missing was "the big picture":

> The study of the principles of Evidence, for a lawyer, falls into two distinct parts. One is Proof in the general sense, the part concerned with the ratiocinative process of contentious persuasion, mind to mind, counsel to Judge or juror, each partisan seeking to move the mind of the tribunal. The other part is Admissibility, the procedural rules devised by the law, based on litigious experience and tradition, to guard the tribunal (particularly the jury) against erroneous persuasion. Hitherto, the latter has loomed largest in our formal studies—has, in fact, monopolized them; while the former, virtually ignored, has been left to the chances of later acquisition, casual and empirical, in the course of practice.
>
> Here we have been wrong; and in two ways:
>
> For one thing, there is, and there must be, a probative science—the principles of proof—independent of the artificial rules of procedure; hence, it can be and should be studied. This science, to be sure, may as yet be imperfectly formulated. But all the more need is there to begin in earnest to investigate and develop it. Furthermore, this process of Proof represents the objective in every judicial investigation. The procedural rules for Admissibility are merely a preliminary aid to the main activity, viz. the persuasion of the tribunal's mind to a correct conclusion by safe materials.
>
> Wigmore, 1937, pp. 3-4

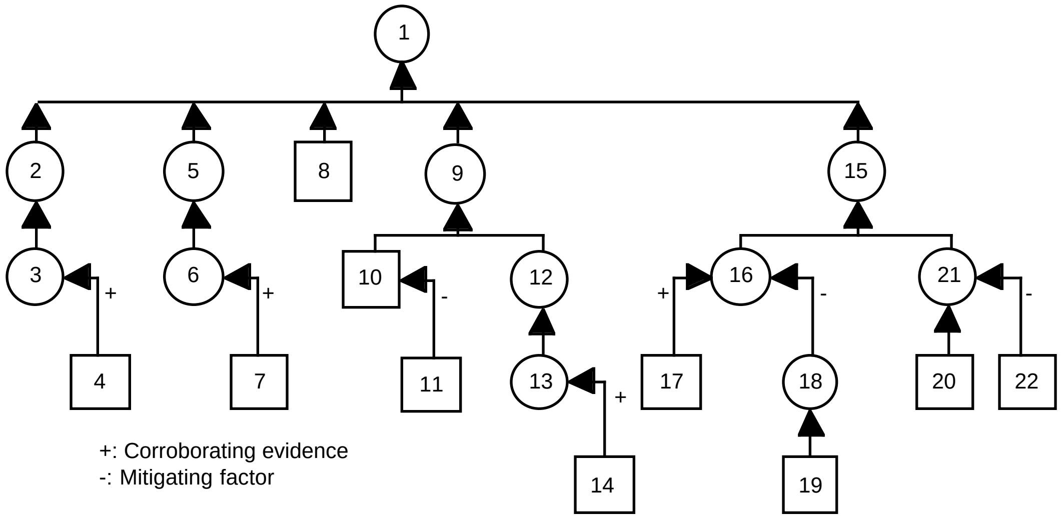
Wigmore thus sought to explicate principles upon which evidence-based inference appeared to be founded in the law. Although every case is unique, he identified recurring patterns in relationships among propositions to be proved (the facta probanda) and propositions that tend to support or refute them (the facta probans). "Basic concepts include conjunction; compound propositions; corroboration; convergence; and catenate inferences (inference upon inference) . . . Each of these notions raises difficult questions about what is involved in determining the overall probative force or weight of evidence" (Twining, 1985, p. 182). To aid understanding of these relationships in particular cases, Wigmore developed a system for charting the structure of arguments. Symbols represent propositions, such as statements of physical evidence, witness testimony, generalizations, or implications of evidence or other propositions; lines among them represent inferential connections. Additional notation, not needed for our purposes, can be used to distinguish among propositions offered by the defense, the prosecution, and the judge, or to suggest the strength and direction of implication.

The *process* of constructing a Wigmore diagram forces careful thought about how evidence leads to inferences and how inferences interrelate, through conjunction, catenation, and so on. This process may be at least as valuable as the product (Twining, 1985, p. 133). The *product*, or the diagram itself, serves to communicate this thinking to others, so that they may be persuaded, or moved to adduce missing themes, counter-explanations, or new lines of evidence to explore. Wigmore's approach can be applied in assessments in which open-ended performances are characterized in terms of established but generally-stated qualities. Just as an apparently simple guilty/not-guilty verdict can be determined by complex arguments from unique data in light of abstract legal principles, a seemingly straightforward numerical rating can involve "questions of what is of value, rather than simple correctness . . . an episode in which students and teachers might learn, through reflection and debate, about the standards of good work and the rules of evidence" (Wolf, Bixby, Glenn, & Gardner, 1991, p. 51).

> **Example 1: Advanced Placement Studio Art Portfolio Assessment**. The purpose of the College Entrance Examination Board's Advanced Placement (AP) Studio Art portfolio assessment is to determine whether high school students exhibit knowledge and skills commensurate with first-year post-secondary art courses (Askin, 1985; Mitchell, 1992). Students develop works for their portfolios during the course of the year, through which they demonstrate the knowledge and skills described in the AP Studio Art materials. The portfolios are rated centrally by artist/educators at the end of the year, using standards set in general terms and monitored by the AP Art advisory committee. At a "standards setting session," the chief faculty consultant and table leaders select portfolios to exemplify the committee's standards. The full team of about 25 readers spends the equivalent of another day of the week-long scoring session examining, discussing, and practicing with these and other examples in order to establish a common framework of meaning. The assessment features 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 guidelines set forth for the sections. 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 mentioned above and a paragraph or two describing the student's goals, intentions, influences, and other factors that help explain the series of works.
>
> Figure 1 is a simplified Wigmore chart based on a discussion of Section B of an Advanced Placement Studio Art portfolio (Myford & Mislevy, 1995). At the top of the diagram is the ultimate probandum, namely, that this

submission should be assigned a rating of 3. Propositions that support or refute this proposition appear below it; propositions that in turn support or refute them appear further below, with the bottom-most propositions closest to the observed data. Several distinct themes appear in the chart. For example, the constellation near the center leading to Proposition #9 concerns the way the project (and the student) developed during the course of the work. The first pieces were weak—evidence which, in and of itself, would tend to move a reader toward a lower rating (#10). But later works, tackling more successfully the same challenge, build strongly from initial efforts (#12). In conjunction, these two propositions support #9, which posits notable progress over time. The constellation at the right leading to #15 concerns evidence about the degree of technical skill exhibited in the work.

Figure 1 also illustrates several catenated or chained inferences, in which propositions play the role of probans, or supporting or refuting evidence, for some inference in the chain, but also play the role of probanda when other propositions are offered in turn to support or refute them. For example, Proposition #3 is evidence about #2, while #3 is itself evidenced by #4. Wigmore noted first that uncertainty accumulates in chained inferences. We would have some degree of uncertainty about the quality of ideation of this project (#2) even if we knew the student had "ingested some difficult art" (#3). We have even more uncertainty about #2 if we do not know #3 directly, but infer it from the references to Jaspar Johns and Lucas Samaras in his written statement (#4)—which may betoken name-dropping rather than knowledge. Wigmore noted secondly that to think through a chain from the bottom up (i.e., inductively), it is useful to consider at each step the weight of evidence offered by the factum probans if it were known to be true: "In dealing with the probative value of the circumstantial class, we are to take the alleged circumstantial . . . fact as somehow believed, then determine its effect. It is immaterial whether it has itself to be proved . . ." (Wigmore, 1937, p. 17). We shall see that this advice is similar in spirit, though opposite in direction, to the way conditional probability structures are used in Pascalian probability-based reasoning with chained inferences.



Key List

1 I agree [with Walter about putting it in the high range—specifically, a rating of 3]

2 . . . if you read the statement, there's a genuine focus on ideation.

3 [We see] a person who has done some, at least been directed to, or has independently gone out and looked at, quite a bit of art that's not easy to ingest and not easy to come to grips with.

4 [The student relates his concentration to the work of Lucas Samaras and Jasper Johns]

5 [We see] the student's involvement as he's working, responding as he's working through the thing.

6 It's pretty obvious that when he's using the material, he really responds to it. He's not just simply opting to do something with the material and then just letting that stay in that point. He does something and seems to maybe see beyond that and through it and say, "Hey, I can do this to it now."

7 I think particularly in the use of the wire [he responds to the material].

8 I think that finding the focus is very strong. He's very much right on track with what he says he's doing.

9 [The pattern of pieces shows development/learning over the course of the work]

10 . . . the beginning elements—the first four of these [would be rated lower].

11 One has to realize, though, I think in the production of art—I think we discussed this some earlier today—about that you're going to have moments where things just don't work.

12 He arranged [the slides] so we would be able to see how he may have evolved through the process.

13 [The later work is] almost unbelievably better than the first works that you see up there . . . the transformation that has occurred on the part of the student is the kind of growth that you would like to see take place in a concentration, rather than being slavish to an idea.

14 . . . something [interesting] is down here [in the later work].

15 [Good, though not excellent, use of materials and formal elements]

16 The only problem that may exist with this is the somewhat looseness of the work

17 It seems to be not as controlled in the sense of skillfully manipulating the materials, in the sense that we traditionally think of it, like if you're directed more toward quote "realistic" work.

18 But I don't have a problem with this [looseness].

19 I find [the looseness] to be very exciting. It's almost kind of, I hate to use the word, but gutsy. The person is obviously one who is very well equipped to taking risks. He's not afraid to really jump into something and really try something rather extraordinary. And I find it to be quite interesting.

20 [There are many close-ups in the submission]

21 There may be some problem maybe in the fact that there are so many close-ups of the work,

22 but I find [the close-ups] to be a way of clarifying to some degree what he's really about in each individual part of the whole unit.

Figure 1. Wigmore chart for rating an AP Studio Art Concentration submission.

The direction of the arrows in a Wigmore diagram indicates a flow of inductive inference. Wigmore was concerned with the difficulty of combining a mass of disparate evidence for ultimate inferences, and he developed his charts to explicate the structure of evidence and inferences. However, he did not claim to prescribe rules for determining that outcome; that is, how to combine a mass of evidence into summary judgments, or to characterize its weight. He left it to the jurors to determine, in a Baconian sense, the extent to which a mass of evidence persuades them of the story of the case. As discussed below, mathematical probability does provide tools for combining evidence within a substantively-determined structure—provided that the crucial elements of the situation can be satisfactorily mapped into the probability framework. The usual problem in jurisprudence is that one would like to know "what really happened," but it is difficult to construct a parameter space comprised of "all the things that could have happened," upon which evidence would induce numerical measures of relative likeliness among all possibilities (i.e., posterior probabilities).

Mathematical Probability

When it is possible to map the salient elements of an inferential problem into the probability framework, powerful tools become available to combine explicitly the evidence that various probans convey about probanda, as to both weight and direction of probative force. Inferential subtleties, such as catenation, missingness, disparateness of sources of evidence, and complexities of interrelationships among probans and probanda, can be resolved. A properly-structured statistical model embodies the salient qualitative patterns in the application at hand and spells out, within that framework, the relationship between conjectures and evidence. It overlays a substantive model for the situation with a model for our knowledge of the situation, so that we may characterize and communicate what we come to believe—as to both content and conviction—and why we believe it—as to our assumptions, our conjectures, our evidence, and the structure of our reasoning.

Perhaps the two most important building blocks of mathematical probability are conditional independence and Bayes theorem. Conditional independence is a tool for mapping Greeno's (1989) "generative principles of the domain" into the framework of mathematical probability, for erecting structures that express the substantive theory upon which deductive reasoning in a field is based.[2] This accomplished, Bayes theorem is a tool for reversing the flow of reasoning—inductively, from observations, through these same structures, to expressions of revised belief about conjectures cast in the more fundamental concepts of the domain, expressed in the language of mathematical probability.

Conditional Independence

Two random variables x and y are *independent* if their joint probability distribution $p(x,y)$ is simply the product of their individual distributions—$p(x,y) = p(x)p(y)$. These variables are unrelated, in the sense that knowing the value of one provides no information about what the value of the other might be. Conditionally independent variables seem to be related—$p(x,y) \neq p(x)p(y)$—but their co-occurrence can be understood as determined by the values of one or more other variables—$p(x,y|z) = p(x|z)p(y|z)$, where the conditional probability distribution $p(x|z)$ is the distribution of x, given the value z of another variable. The conjunction of sneezing, watery eyes, and a runny nose described as a "histemic reaction" could be triggered by various causes such as an allergy or a cold; the specific symptoms play the role of x's and y's, while the status of "having a histemic reaction" plays the role of z. The paradigms of a field supply

[2] Conditional independence also plays a key role in justifying the use of mathematical probability-based reasoning for real-world problems. The layman unfamiliar with probability and statistics, other than through informal notions about random sampling and large samples, might question whether mathematical probability has anything to do with real-world observations that are governed by disparate mechanisms and may be linked with one another in unknown ways (e.g., prospective test scores of students about whom we know nothing other than that each surely brings a unique personality and history to the tasks, aspects of which are similar to certain other students in some ways but not in other ways). Even if we admit the possibility, indeed the inevitability, of such differences among the antecedents of potential observations, yet at a given point in time have no information to distinguish among them a priori, then these observations are "exchangeable" from our point of view. That is, our subjective probability distribution for their scores would be the same under any permutation of the variables. Even if the mechanism by which values are produced is nothing like random, de Finetti's Theorem (de Finetti, 1974) says the distribution of finite subsets of an infinite sequence of exchangeable variables can be expressed as the expectation, over a mixing distribution, of conditionally independent and identically distributed (iid) variables. Diaconis and Freedman (1980) show further that conditionally iid representations can be used to approximate subsets of finite sets of exchangeable variables, with increasing fidelity for larger sets. Thus, the use of mathematical probability need not be justified by the manner in which values of variables arise, but by our state of knowledge about them. Of course if we learn more about influences and mechanisms that produce values of variables, we can improve our model of the situation. Variables that were exchangeable in light of previous knowledge need not be later. The interested reader is referred to Lindley and Novick (1981) for an exploration of the role of exchangeability vis-à-vis random sampling and populations in connection with inference in experimental and nonexperimental settings.

"explanations" of phenomena in terms of concepts, variables, and putative conditional independence relationships. Judah Pearl (1988) argues that inventing intervening variables is not merely a technical convenience, but a natural element in human reasoning:

> Conditional 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. (p. 44*)*

In psychology, Charles Spearman's *methodological* insight was that conditional independence of observable scores in standardized tests, given an unobservable "intelligence" variable *g*, would imply particular patterns of relationships among the observable scores (Spearman, 1904, 1927). Now while conditional independence is thus used to express Spearman's psychological concept of a trait that determines behavior across a broad array of situations, the mathematical concept of conditional independence *per se* in no way implies *g* or anything like it. Indeed, Examples 4 and 5 below show how conditional independence is used to express psychological theories under which the interactions between persons' knowledge structures and the situations they encounter are central to understanding behavior. The point is that Spearman's inferential machinery, as distinct from his psychological theory, supplied a framework for reasoning deductively and inductively within his paradigm, and, at least in principle, for disconfirming conjectures about behavior in terms of hypothesized traits.

The tradition of statistical inference founded upon unobservable variables and induced conditional probability relationships now dominant in educational and psychological measurement thus extends back to Spearman's early work, bolstered by Wright's (1934) path analysis, Lazarsfeld's (1950) latent class models, and more recent work on structural equation modeling in the presence of measurement errors (e.g., Jöreskog & Sörbom, 1979). Lewis (1986) notes continued and considerable extensions of the logic of inference for problems involving unobservable variables, exploring possibilities and limitations, developing statistical machinery for estimation and prediction (e.g., Holland & Rosenbaum, 1986; Rasch, 1960/1980). The first part of Example 2 (below) illustrates how deductive reasoning flows from the conditional probability relationship at the core of Rasch's (1960/1980) item response theory (IRT) model for dichotomous test items.

> **Example 2: An Item Response Theory Model**. The Rasch model for dichotomous test items is used to structure inference about students' overall level of proficiency in a specified domain of test items. It posits that responses to *n* test items from the domain are conditionally independent, given parameters characterizing a student's overall tendency to make correct responses (denoted θ) and each item's difficulty (β_j denoting the difficulty parameter for Item *j*):
>
> $$\mathrm{P}\left(x_1,\ldots,x_n|\theta,\beta_1,\ldots,\beta_n\right)=\prod_{j=1}^{n}\mathrm{P}\left(x_j|\theta,\beta_j\right), \tag{1}$$
>
> with
>
> $$\mathrm{P}\left(x_j|\theta,\beta_j\right)=\frac{\exp\left[x_j\left(\theta-\beta_j\right)\right]}{\left[1+\exp\left(\theta-\beta_j\right)\right]}, \tag{2}$$
>
> where x_j is the response to Item *j* (1 for right, 0 for wrong). Figure 2 shows the probabilities of correct response to three items, with difficulty parameters -1, 0, and +1, as a function of θ. Low values of θ indicate lower chances of correct response and high values indicate higher chances, at rates determined by the item parameters.
>
> Figure 3 depicts the relationships expressed in (1) among the variables pertaining to a single student as a directed acyclic graph (DAG). Each node represents a variable—one proficiency variable, θ, and three items, x_1, x_2, and x_3. An arrow between nodes represents a conditional probability relationship between variables, the direction signifying which variable is being conditioned on (from "parents" to "children," in DAG terminology from genetic applications). The lack of arrows among the individual *x*'s represents the conditional independence indicated in (1); they are posited to be unrelated except through θ. For any given x_j, the probability distribution is modeled as

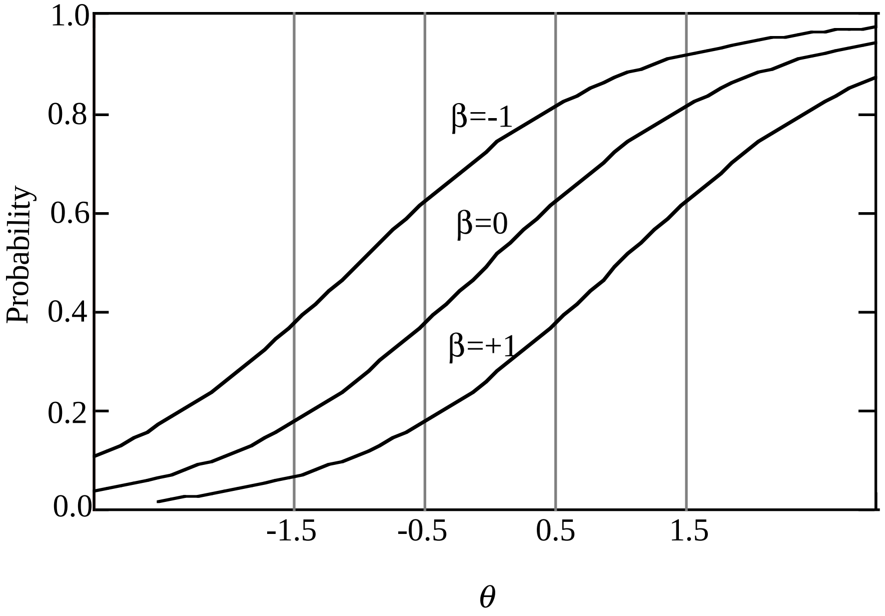


Figure 2. Probability of a correct response, conditional on θ, for items with β = -1, =0, and =1.

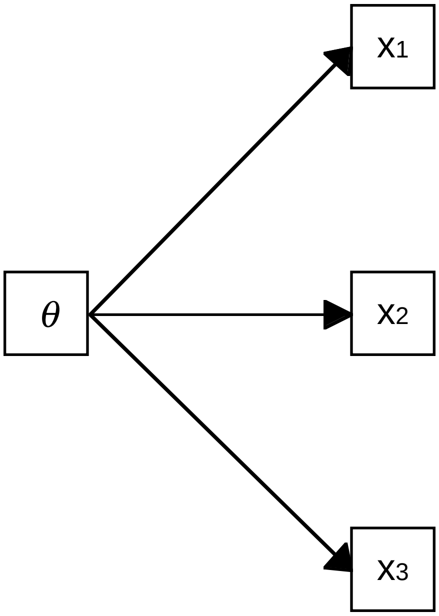


Figure 3. Directed acyclic graph for the IRT example.

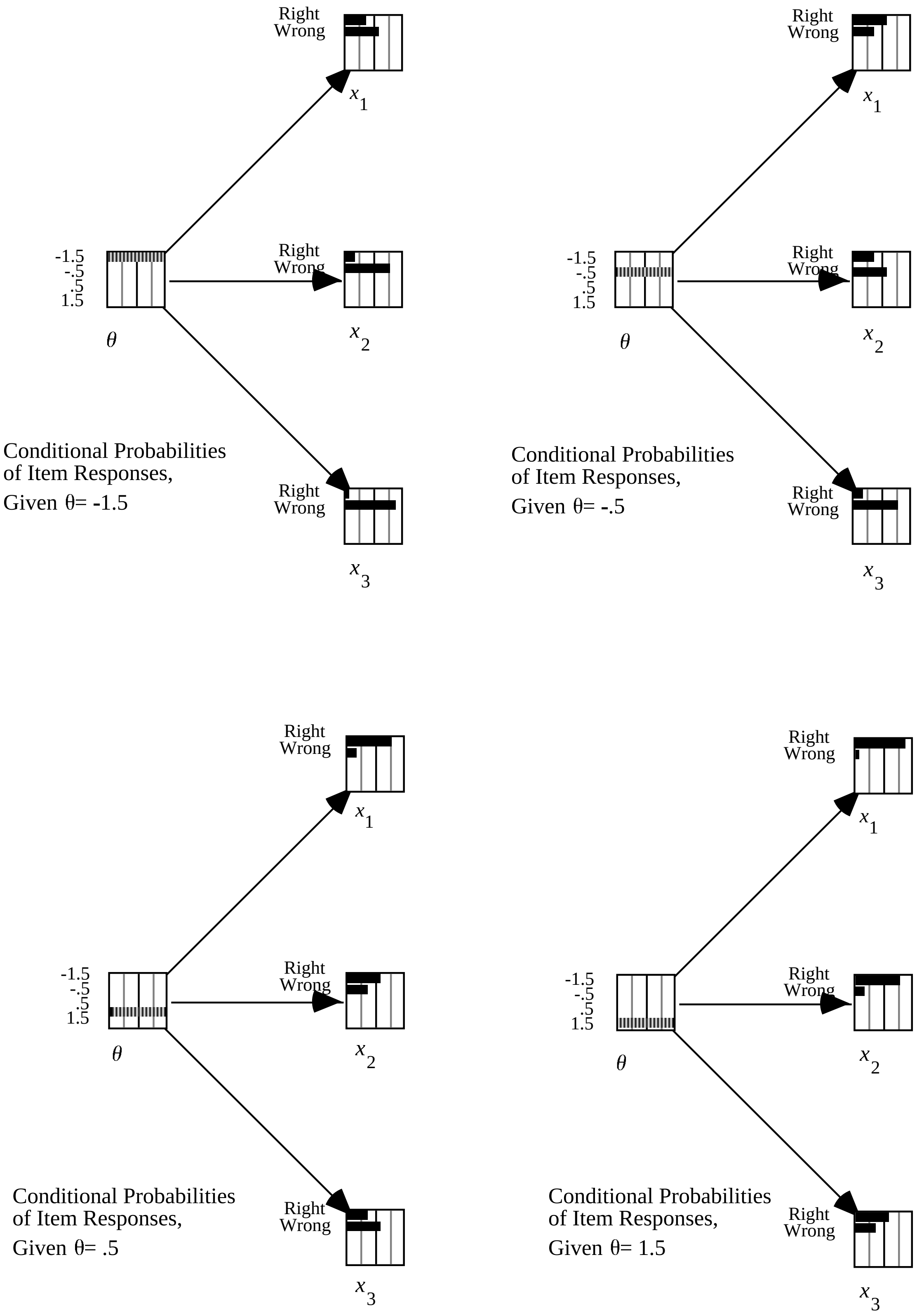
depending on θ and β_j as indicated in (2). Equations (1) and (2) represent deductive inference from θ and β's to expectations about x's; that is, θ and β's are probans, the x's, probanda. Alternatively stated, if particular values of θ and β's were given, we could use (1) and (2) to assign probabilities, or numerical statements of our expectations, to conjectures about observable responses such as "The response to Item 1 will be 0 rather than 1" or "All three responses will be correct as opposed to a pattern with at least one 0."

The arrows in Figure 3 indicate the structure of relationships, but not their strengths. Suppose for simplicity that θ can take only four values, -1.5, -.5, .5, and 1.5, and we know the β values of the three items to be -1, 0, and 1 respectively. Table 1 gives the probabilities of correct response to each of the items conditional on each possible θ value, as calculated from (3). These relationships are depicted as augmented DAGs in the four panels of Figure 4. Each panel depicts the probabilities of right and wrong item responses if θ is known with certainty to take one of its four possible values. Bars in the nodes corresponding to items represent probabilities from Table 1 for right and wrong responses, given the θ values. The bar for the θ node goes all the way to 1 for the keyed θ value in each panel, thereby conditioning expectations for x's that would follow (deductively) if it were the true value.

Table 1

Conditional Probabilities of Correct Response in IRT Example

	Item parameter (β)		
Student parameter (θ)	-1	0	1
-1.5	.378	.182	.076
-.5	.622	.378	.182
.5	.818	.622	.378
1.5	.924	.818	.622



Nodes represent variables; bars represent probabilities of potent values of a variable, summing to one, with a dashed bar to one representing certainty.

Figure 4. Probabilities of item responses, given student proficiency.

Bayes Theorem

We must reason inductively in most practical applications. In the IRT example, we observe item responses x in order to increase our knowledge about a student's level of proficiency on tasks in the domain. If we know or have good estimates of the β's, then the x's are now probans and θ the probandum. That is, given a particular pattern of item responses, we wish to express our belief about conjectures about θ such as "θ=-1.5." Since we can map the possibilities into the probability framework in this case, Bayes theorem provides a mechanism for accomplishing the desired inductive inference.

In general terms, let x be a variable whose probability distribution $p(x \mid z)$ depends on the variable z. Suppose also that prior to observing x, belief about the value of z can be expressed in terms of a probability distribution $p(z)$ For example, we may consider all possible values of z equally likely, or we may have an empirical distribution based on values observed in the past. Bayes theorem says

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}, \tag{3}$$

where $p(x)$ is the expected value of x over all possible values of z, or

$$p(x) = E[p(x|z)] = \begin{cases} \int p(x|z)p(z)d(z) & z \text{ continuous} \\ \sum p(x|z)p(z) & z \text{ discrete} \end{cases}, \tag{4}$$

with the integral or sum taken over the admissible range of z (Box & Tiao, 1973, p. 10).

We see in (3) that the terms that change belief about a conjecture, from $p(z)$ to $p(z \mid x)$, are the so-called likelihoods, $p(x \mid z)$; that is, the relative probabilities of the observed datum given each of the possible states that might have produced it. While the expressions $p(x|z)$ drive *deductive* reasoning about possible x's for a given z, the same expressions drive *inductive* reasoning about the likelihood of possible z's once a particular value of x is observed. If, for a particular value of x, $p(x|z_1)$ is twice $p(x|z_2)$, then observing this value of x argues in and of itself twice as strongly for z_1 as for z_2, independently of our prior state of belief about their relative prospects and of evidence from other sources (this latter information to be taken into account in ways discussed below in connection with inference networks). From a Bayesian statistical perspective, likelihoods characterize completely the weight and direction of evidential value that observations bear for a conjecture.

This last point deserves emphasis, for it is the essence, the characterization of belief and weight of evidence under the paradigm of mathematical probability:

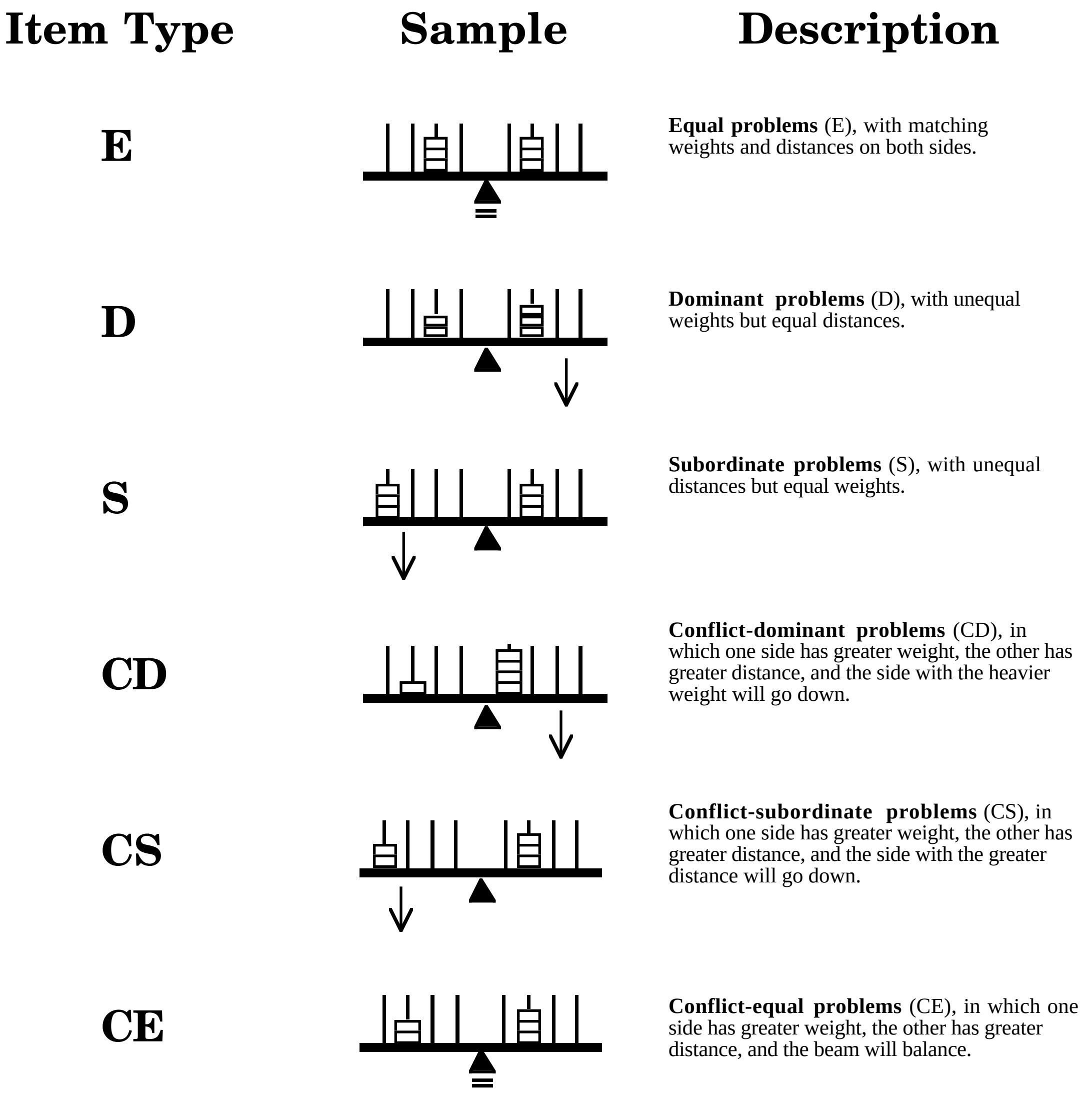
- Prior to observing a datum, relative belief in a space of possible propositions is effected as a probability (density) distribution, namely, the prior distribution $p(z)$.

- Posterior to observing the datum x, relative belief in the same space is effected as another probability (density) distribution, the posterior distribution $p(z \mid x)$.

- The evidential value of the datum x is conveyed by the multiplicative factor that revises the prior to the posterior for all possible values of z, namely, the likelihood function $p(x \mid z)$. One examines the *direction* by which beliefs associated with any given z change in response to observing x (is a particular value of z now considered more probable or less probable than before?) and the *extent* to which they change (by a little or by a lot?).

Example 3: A Latent Class Model. "Achievement testing as we have defined it is a method of indexing stages of competence through indicators of the level of development of knowledge, skill, and cognitive process," submitted Glaser, Lesgold, and Lajoie (1987, p. 81); "These indicators display stages of performance that have been attained and on which further learning can proceed." The important questions for guiding learning are not "How many items did this student answer correctly?" or "What proportion of the population would have scores lower than his?" but, in Thompson's (1982) words, "What can this person be thinking so that his actions make sense from his perspective?" and "What organization does the student have in mind so that his actions seem, to him, to form a coherent pattern?" This example shows how a series of tasks devised by Robert Siegler (1981) and a latent class statistical model (Lazarsfeld, 1950) support probability-based inference about such aspects of children's proportional reasoning as viewed from the perspective of a neo-Piagetian paradigm (also see Kempf, 1983).

Jean Piaget proposed that children develop proportional reasoning in stages that reflect increasing awareness of the salient properties of a problem class, and increasing sophistication in how they combine to produce a solution (Inhelder & Piaget, 1958). Conjectures about children's proficiency under Piaget's developmental paradigm concern the stages of development at which they are functioning, and observable data consist of their words and actions as they solve proportional reasoning tasks. Siegler's tasks show varying numbers of weights placed at varying locations on a balance beam, and a child predicts whether the beam will tip to the left, tip to the right, or remain in balance. The six basic types of task are illustrated in Figure 5. Following Piaget, Siegler hypothesized that children could be classified into one of five stages: four characterized by how many of the cumulative reasoning rules shown in Table 2 they had acquired—representing Stages I through IV—and an earlier "pre-operational" Stage 0 in which neither weight nor distance from the fulcrum are seen to bear any systematic relationship to the movement of the beam.



Item Type	Sample	Description
E		**Equal problems** (E), with matching weights and distances on both sides.
D		**Dominant problems** (D), with unequal weights but equal distances.
S		**Subordinate problems** (S), with unequal distances but equal weights.
CD		**Conflict-dominant problems** (CD), in which one side has greater weight, the other has greater distance, and the side with the heavier weight will go down.
CS		**Conflict-subordinate problems** (CS), in which one side has greater weight, the other has greater distance, and the side with the greater distance will go down.
CE		**Conflict-equal problems** (CE), in which one side has greater weight, the other has greater distance, and the beam will balance.

Figure 5. Basic types of balance beam tasks.

Table 2

Successive Rules Children Are Posited to Acquire as Proportional Reasoning Develops

Rule I:	**If the weights on both sides are equal, the beam will balance. If they are not equal, the side with the heavier weight will go down.** Weight is the "dominant dimension" in this domain of tasks, because children are generally aware that weight is important in the problem earlier than they realize that distance from the fulcrum, the "subordinate dimension," also matters.
Rule II:	**If the weights and distances on both sides are equal, then the beam will balance. If the weights are equal but the distances are not, the side with the longer distance will go down. Otherwise, the side with the heavier weight will go down.** A child using this rule uses the subordinate dimension only when information from the dominant dimension is equivocal.
Rule III:	**Same as Rule II, except that if the values of both weight and distance are unequal on both sides, the child will "muddle through"** (Siegler, 1981, p. 6). A child using this rule now knows that both dimensions matter, but doesn't know just how they combine.
Rule IV:	**Combine weights and distances correctly** (i.e., compare torques, or products of weights and distances).

Note. These rules are based on Seigler's (1981) presentation. Stage x signifies being able to apply all rules up through and including Rule x.

If the underlying developmental theory were perfect, children's stages of reasoning would tightly control the rates at which they would respond correctly to the various types of tasks; these rates are shown as Table 3. But because the model is not perfect,[3] and because children make slips and

[3] This model assumes that there are five exhaustive and mutually exclusive states. Alternative models could be used to relax these restrictions. The section on abductive reasoning discusses the role of detecting unexpected response patterns for tempering inference in specific cases, and for gaining insights on how to refine or revise a provisional model.

Table 3

Theoretical Conditional Probabilities of Correct Response in Balance Beam Example

	Task type					
Stage	E	D	S	CD	CS	CE
0	.333	.333	.333	.333	.333	.333
I	1.000	1.000	.000	1.000	.000	.000
II	1.000	1.000	1.000	1.000	.000	.000
III	1.000	1.000	1.000	.333	.333	.333
IV	1.000	1.000	1.000	1.000	1.000	1.000

lucky guesses, any response could be observed from a child in any stage. A latent class model can be used to express the expectations of correctness of the various tasks at each of the stages, while allowing for some "noise" in real data (Mislevy, Yamamoto, & Anacker, 1992). Instead of positing that children in Stage II will with certainty respond incorrectly to "Conflict-Dominant" tasks, we might instead estimate the proportion of correct answers, or P(CD=correct | Stage=II). These probabilities play the same role as the item parameters β in the IRT example, quantifying expectations of potential observations x (in this case, predictions about which way the balance beam will move) given the unobservable psychological variable of interest θ (in this case, the child's stage of reasoning). Estimated values for proportions of correct response given reasoning stages appear in Table 4.

Table 4

Estimated Conditional Probabilities of Correct Response in Balance Beam Example

	Task type					
Stage	E	D	S	CD	CS	CE
0	.333[a]	.333[a]	.333[a]	.333[a]	.333[a]	.333[a]
I	.973	.973	.026	.973	.026	.026
II	.883	.883	.883	.883	.116	.116
III	.981	.981	.981	.333[a]	.333[a]	.333[a]
IV	.943	.943	.943	.943	.943	.943

[a] Denotes fixed value for estimation. Note also that true-positive and false-positive probabilities within a given stage are constrained to be equal across task types, to ensure that the latent class model is identified.

A child in Stage I usually predicts the side with more weight will go down, although different distances from the fulcrum may cause the other side to go down or the beam to remain in balance; it is necessary to compare torques to know. But in CD tasks the side with more weight actually does go down, and the Stage I child gets the right answer for the wrong reason! When a child's understanding deepens to the point at which he realizes distance matters but doesn't know how to combine it with weight, he is less likely to get CD tasks right than when he was in Stage I. Because probabilities of correct response to CD tasks do not increase monotonically with increasing total test scores, they provide weak evidence for the inferential problem IRT is meant to address, namely gauging overall tendency to make correct responses. From the perspective of the developmental theory, however, not only is this reversal expected, it provides useful evidence for distinguishing among children with different ways of thinking about the domain. Succeeding with the more complex "Conflict-Dominant" (CD) tasks while missing the simpler "Subordinate" (S) tasks is converging evidence that a child is reasoning in Stage I. This pattern highlights the distinction between Wigmore's two terms, "corroborating evidence" and "converging evidence." Corroborating evidence refers to repeated, consonant observations of the same kind of data for the same conjecture: Consistently correct CD responses are corroborating evidence for inferring proficiency in the subdomain of CD tasks; consistently incorrect S responses are corroborating evidence for inferring proficiency in the subdomain of S tasks. Converging evidence refers to patterns of data of different kinds that are consistent with a conjecture: Correct CD responses together with incorrect S responses are converging evidence about membership in Stage I.

Though cast within a different psychological paradigm, the DAG for this model is similar in structure to that of the Rasch model: A single unobservable variable (stage of reasoning) is posited to determine probabilities of task outcomes (correct and incorrect predictions about the balance-beam movement). Suppose that our beliefs that a student is in each of the stages from 0 through IV before we observe a response to any task, corresponding to the values of $p(z)$ in the expression above for Bayes theorem, are given by the Mislevy et al. estimates of proportions of children at each of the stages in Siegler's sample:

$$\begin{aligned}&(\mathrm{P}(\mathrm{Stage}=0),\mathrm{P}(\mathrm{Stage}=\mathrm{I}),\mathrm{P}(\mathrm{Stage}=\mathrm{II}),\mathrm{P}(\mathrm{Stage}=\mathrm{III}),\mathrm{P}(\mathrm{Stage}=\mathrm{IV}))\\&\qquad=(.257,.227,.163,.275,.078).\end{aligned}$$

This state of knowledge is depicted in the first panel of Figure 6, showing for simplicity only the nodes for Stage Membership and one task of each type. Suppose now we observe a correct response to a S task. The values in the S column of Table 4 correspond to the values of $p(x \mid z)$ with x being "correct response to a S item" and with z taking the values of the five possible stage memberships. These values register the evidential value of a correct-S observation with respect to inference about a student's stage of understanding, shifting belief upwards in general, and away from Stage I and toward Stage III in particular. Updated beliefs about a student's stage membership, or values of $p(z \mid x)$ with x and z interpreted as above, are then obtained in two steps, through first (4) then (3) as follows:

$$\begin{aligned}
\mathrm{P}(x) &\equiv \mathrm{P}(\text{Correct response to S}) \\
&= \sum_{j=1}^{5} \mathrm{P}(\text{Correct response to S}|\text{Stage} = j)\mathrm{P}(\text{Stage} = j) \\
&= (.333)(.257) + (.026)(.227) + (.883)(.163) + (.981)(.275) + (.943)(.078) \\
&= .086 + .006 + .144 + .270 + .073 = .579.
\end{aligned}$$

$$\begin{aligned}
&(\mathrm{P}(\text{Stage} = 0|\text{Correct response to S}), \ldots, \mathrm{P}(\text{Stage} = \text{IV}|\text{Correct response to S})) \\
&\qquad = \left(\frac{.086}{.579}, \frac{.006}{.579}, \frac{.144}{.579}, \frac{.270}{.579}, \frac{.073}{.579}\right) \\
&\qquad = (.149, .010, .249, .466, .126).
\end{aligned}$$

These revised beliefs, as well as updated expectations for possible future responses to other task types, appear in the second panel of Figure 6.

The keys to successful exploitation of probability-based reasoning in a given application are the definitions of variables to capture the salient elements of the situation, and the structuring of probability distributions and conditional independences that capture the most important relationships among those elements. It may be painstaking and difficult work to model subtleties of the kinds mentioned above (see, for example, how Schum, 1981, sorted out intricacies of witness credibility), and it may be necessary to add additional layers of parameters to express uncertainty about relationships. Nevertheless, if the relationships necessary for deductive reasoning and prior beliefs about unknown

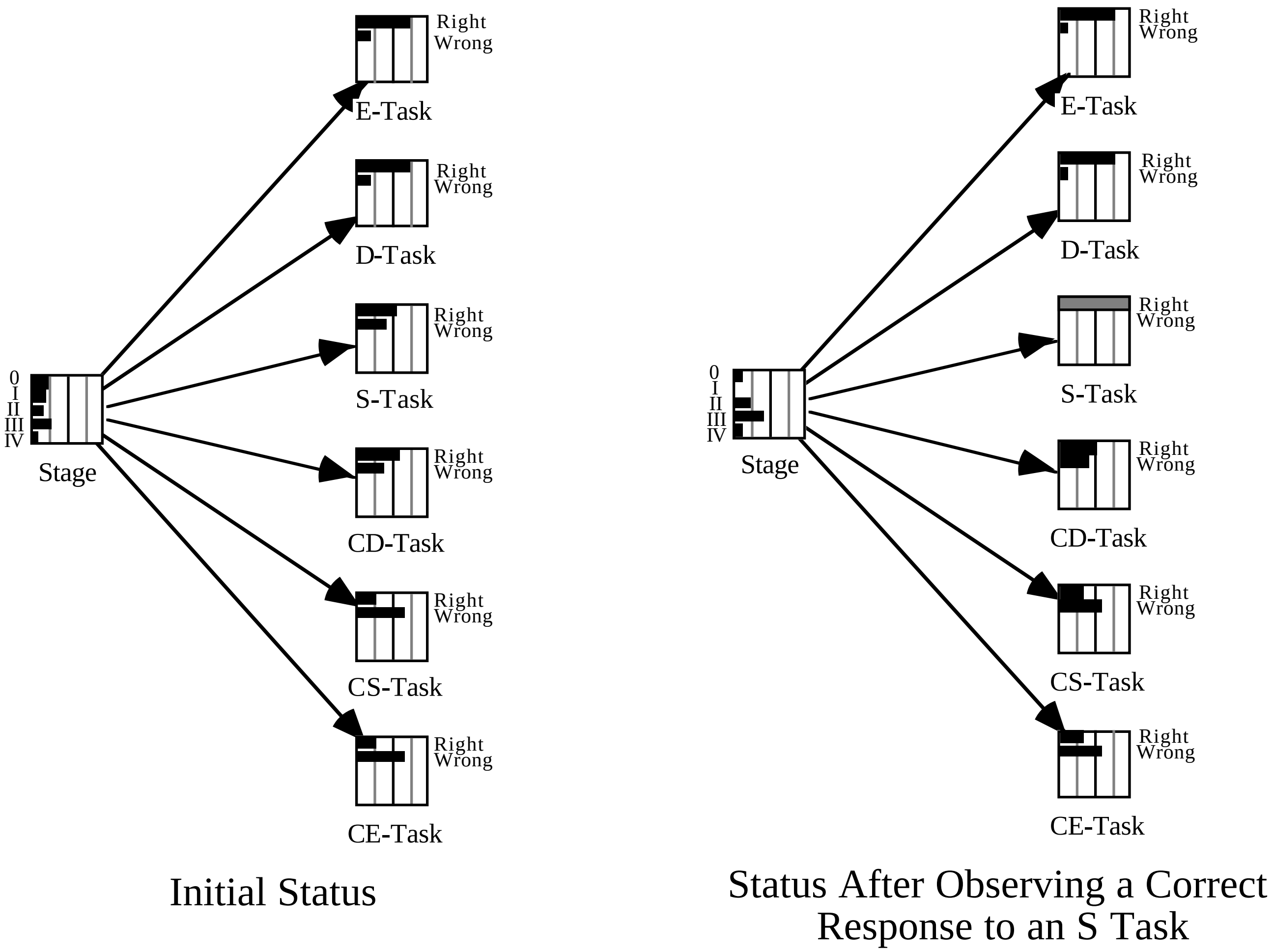


Figure 6. Belief in balance beam example, before and after observing a correct response to an S task.

parameters can be mapped into the framework of mathematical probability, then Bayes theorem can provide principled inductive reasoning that accounts for the subtleties within the same framework.

Bayesian Inference Networks

Applying Bayes theorem in its textbook form (Equations 3 and 4) becomes unwieldy rather quickly as the number of variables in a problem increases. Efficient probability-based inference in complex networks of interdependent variables is an active 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). Interest centers on obtaining the distributions of selected variables conditional on observed values of other variables, such as likely characteristics of offspring of selected animals given characteristics of their ancestors, or probabilities of disease states given symptoms and test results. The conditional independence relationships suggested by substantive theory play a central role in the topology of the network of interrelationships in a system of variables. If the topology is favorable, such calculations can be carried out efficiently through extended application of Bayes theorem even in very large systems, by means of strictly local operations on small subsets of interrelated variables ("cliques") and their intersections. Discussions of construction and local computation in Bayesian inference networks can be found in the statistical and expert-systems literature (see, for example, Lauritzen & Spiegelhalter, 1988, Pearl, 1988, and Shafer & Shenoy, 1988; computer programs that carry out the required computations include Andersen, Jensen, Olesen, & Jensen, 1989, and Noetic Systems, 1991).

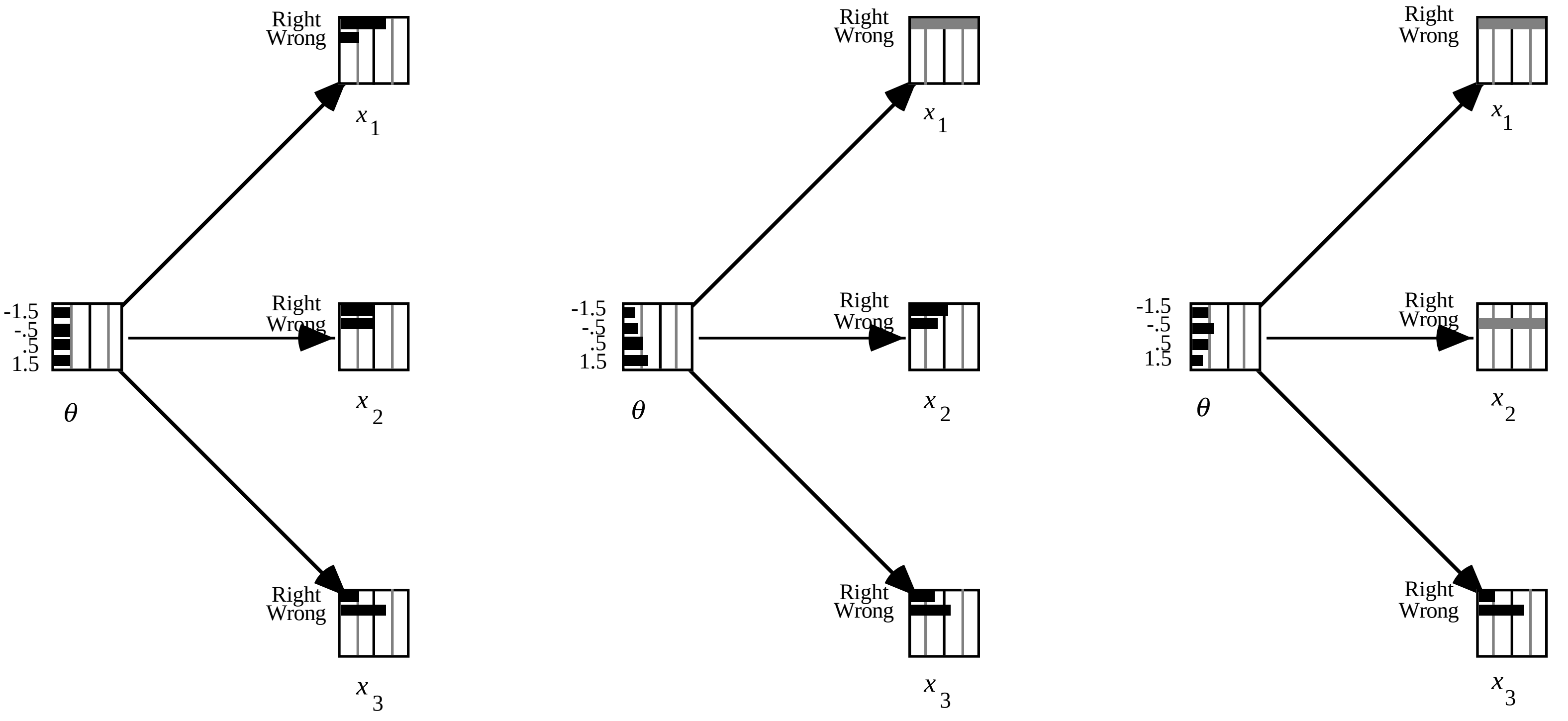
A *recursive representation* of the joint distribution of a set of random variables $x_1,\ldots,x_N$ takes the form

$$p(x_1,\ldots,x_n) = p(x_n|x_{n-1},\ldots,x_1)p(x_{n-1}|x_{n-2},\ldots,x_1)\cdots p(x_2|x_1)p(x_1)$$

$$= \prod_{j=1}^{n} p(x_j|x_{j-1},\ldots,x_1), \tag{5}$$

where the term for j=1 is defined as simply $p(x_1)$. A recursive representation can be written for any ordering of the variables, but one that exploits conditional independence relationships is more useful because variables drop out of the conditioning lists. This is equivalent to omitting arrows ("edges") from the DAG, thus simplifying the topology of the network. It is here that substantive theory comes into play, in (a) defining unobservable variables that characterize students' state or structure of understanding, and observable variables that will convey evidence about that understanding, and (b) defining intervening variables and conditional independences through which deductive reasoning flows, so as to capture important substantive relationships and simplify computations. An inference network for medical diagnosis, for example, includes nodes for symptoms and test results, which are observable, and for syndrome and disease states, which are not observable, but in terms of which theories of the progression and treatment of disease are framed (Andreassen, Jensen, & Olesen, 1990). Analogously, an inference network for cognitive diagnosis includes nodes for students' actions and explanations and conditions of assessment situations, which are observable, and for skill and knowledge states, which are not, but in terms of which theories of knowledge and learning are framed (Martin & VanLehn, 1993; Mislevy, 1995).

> **Example 2: An IRT Model, continued**. This section extends the IRT example to sequential gathering and evaluating of evidence, or adaptive testing (Wainer et al., 1990), and uncertainty about item parameter values—still with examinees' overall proficiency the target of inference. Suppose that prior belief about an examinee's θ, before seeing any item responses, is characterized by equal probabilities of .25 for each of the four possible values posited above. (Alternatively, prior beliefs might be based on his results from earlier tests, empirical distributions of other examinees who have been tested, or on knowledge of his instructional history.) Assuming the probabilities of correct response given in Table 4 conditional on each possible θ, we can deduce probabilities that represent our expectations of seeing correct responses from a student about whom we have no additional information. These are depicted in the first panel of Figure 7. If we now observe a correct response to Item 1, we can apply Bayes theorem to update our beliefs about this examinee's θ, as shown in the second panel. But once our belief about θ is revised through inductive reasoning from x_1, we reason deductively to update our expectations for Items 2 and 3. The second panel of Figure 7 thus shows (a) certain knowledge about the response to Item 1, (b) a shift of belief about θ to higher values, and (c) greater expectations of correct response to the items not yet presented. The third panel shows the results of another cycle of inductive reasoning (from observing x_2 to belief about θ) followed by deductive reasoning (from revised belief about θ to revised expectations about x_3), that are initiated by an incorrect response to Item 2.
>
> Figures 4 and 7 treat as known the conditional probabilities for x's given θ implied by item parameters β and the prior distribution $p(\theta)$; only uncertainty concerning an individual student's θ and x's is addressed. This may be reasonable when strong evidence is available about these quantities, but in principle they too are never known with certainty. We learn something about them inductively from responses of several students to several items. A more complete Bayesian treatment of the IRT setup includes unknown parameters τ for the distribution of θ, parameters ξ for the distribution of β's, and hyperparameters η and ζ for the distributions of τ and ξ (Mislevy, 1986; this setup can be further extended to incorporate information from



Initial State of Belief

State of Belief after Observing a
Correct Response to Item 1

State of Belief after Observing a
Correct Response to Item 1 and an
Incorrect Response to Item 2

Nodes represent variables; bars represent probabilities of potent values of a variable, summing to one, with a dashed bar to one representing certainty.

Figure 7. Posterior probabilities for proficiency after observing a sequence of item responses.

collateral information about students, as in Mislevy and Sheehan, 1989, and collateral information about tasks, as in Mislevy, Sheehan, and Wingersky, 1993). As a particular instance of (5), we might thus posit

$$p(x,\theta,\beta,\xi,\tau,\eta,\zeta) = p(x|\theta,\beta)p(\theta|\tau)p(\tau|\eta)p(\eta)p(\beta|\xi)p(\xi|\zeta)p(\zeta),$$

and, after observing only response vectors *x* from a collection of students to a collection of tasks, calculate approximate posterior distributions for any item or population parameters of interest, or for task or individual student parameters taking uncertainty about higher level parameters into account. A portion of a corresponding extended DAG appears as Figure 8.

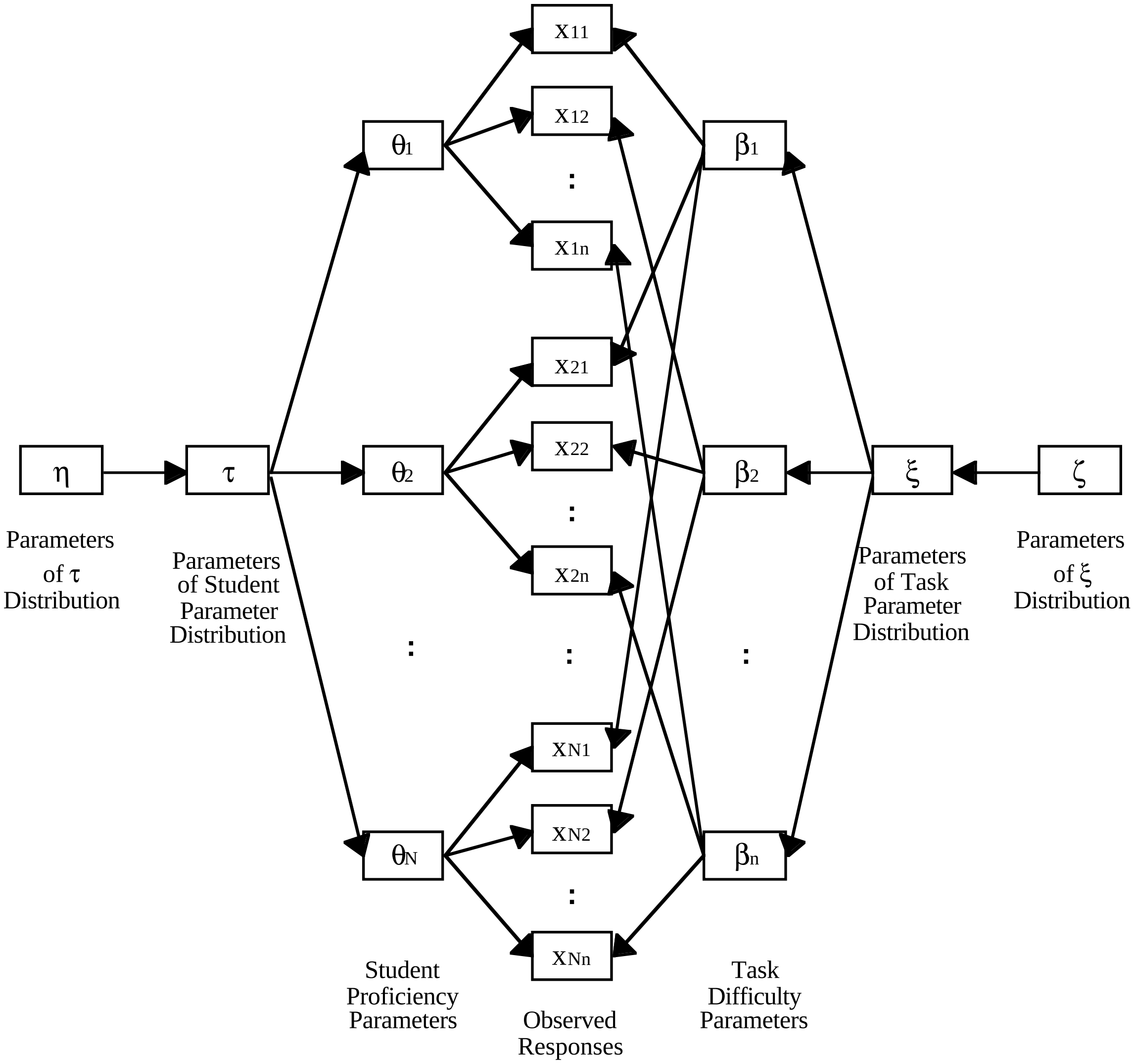

η
τ
θ1
θ2
:
θN
X11
X12
:
X1n
X21
X22
:
X2n
:
XN1
XN2
:
XNn
β1
β2
:
βn
ξ
ζ
Parameters of τ Distribution
Parameters of Student Parameter Distribution
Student Proficiency Parameters
Observed Responses
Task Difficulty Parameters
Parameters of Task Parameter Distribution
Parameters of ξ Distribution

Figure 8. More complete directed acyclic graph for the IRT example.

Example 4: Mixed-Number Subtraction. The data in this example are again familiar right/wrong responses to open-ended mixed-number subtraction problems, but inference now concerns a more complex student model meant to support short-term instructional guidance. We see how conditional independence relationships can structure and support inference for a psychological model under which the difficulty of an item depends on the strategy a student employs—a source of uncertainty for inferences about overall proficiency, but a source of evidence for inferences about strategy usage. We further see how the interrelationships among skills and between skills and observable responses exemplify some of Wigmore's basic evidential structures, and how they are handled in the framework of mathematical probability. The data and the cognitive model are due to Tatsuoka (1987, 1990). The 530 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.

Mislevy (1995) characterizes 15 items in terms of which of seven subprocedures are required to solve each item with Method A and with Method B. The corresponding student model consists of a variable for which strategy a student characteristically uses, and which of the seven subprocedures the student is able to apply. The structure connecting the observable responses to the unobservable student-model parameters is that ideally, a student using, say, Method A would correctly answer items that under that strategy require only subprocedures the student has at his disposal (Falmagne, 1989; Haertel & Wiley, 1993; Tatsuoka, 1990). But 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).

Figure 9 depicts an inference network for Method B only. Five nodes represent basic subprocedures that a student who uses Method B needs to solve various kinds of items; these are labeled Skill 1 through Skill 5. Conjunction, one of the basic evidential structures described by Wigmore, appears in this DAG: The conjunctive node "Skills1&2," for example, takes the value "yes" if and only if a student has both Skill 1 and Skill 2. Each node for the observable response to a particular subtraction item is the child of a node representing the minimal conjunction of skills needed to solve it with

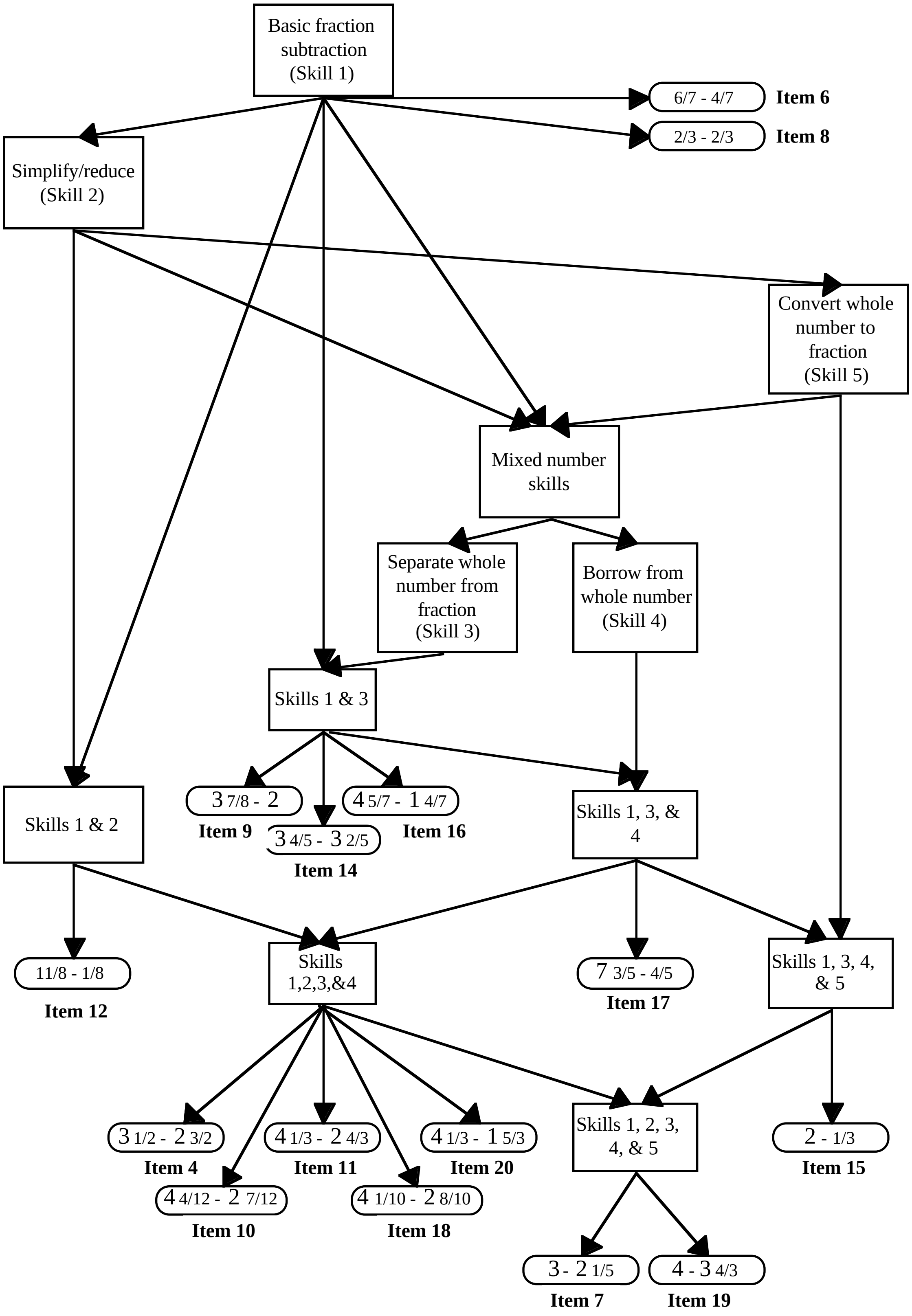


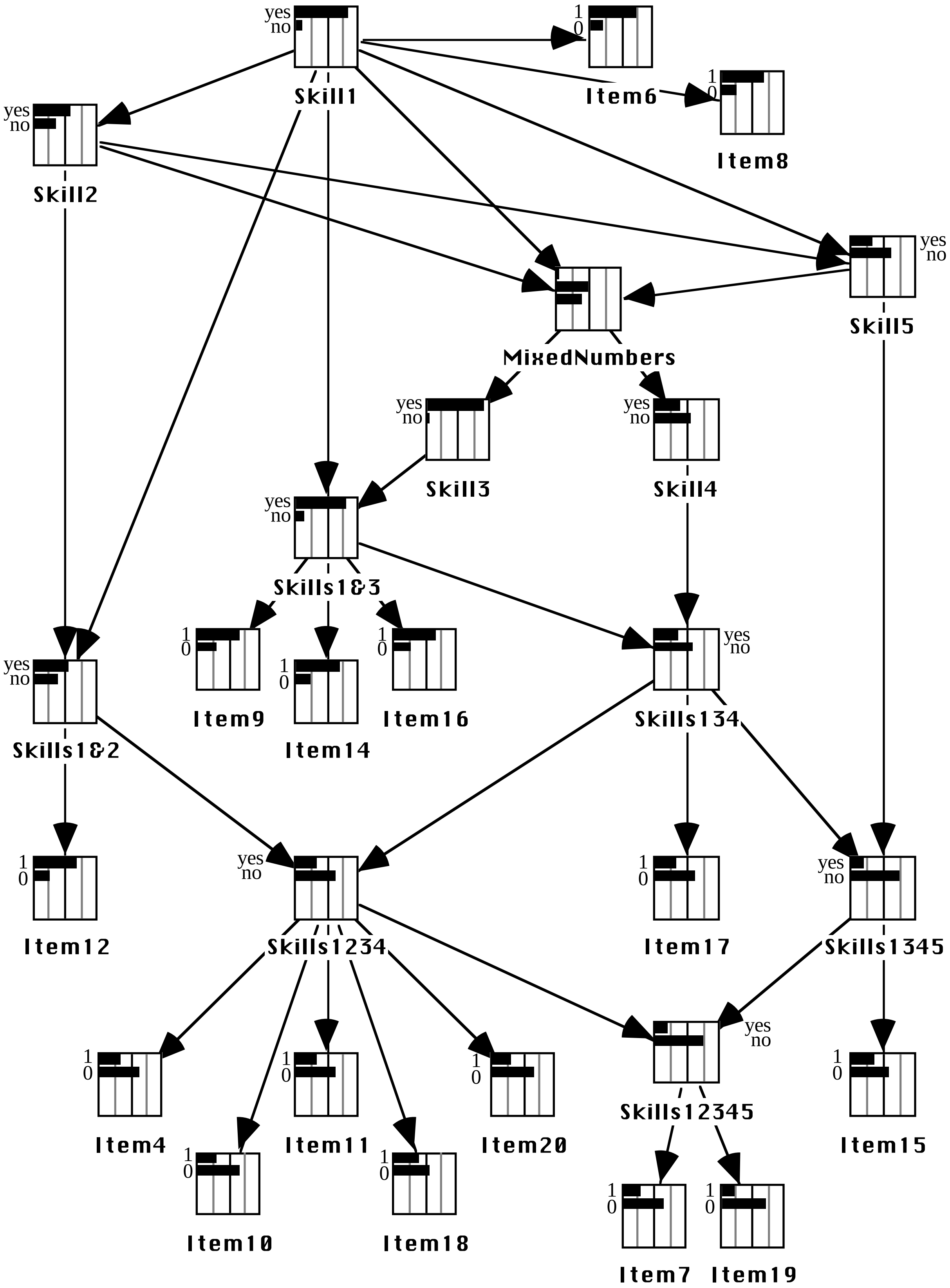
Figure 9. Directed acyclic graph for Method B.

Method B. The relationship between such a node and an item incorporates false positive and false negative probabilities. Catenation, another of Wigmore's basic structures, appears in chains such as the one from "Skill 2" to "Skills1&2" to "Item 12." Inference in this chain is structured through the conditional probability distributions of Item 12 responses given each possible value of "Skills1&2" as if it were true, and the conditional probability distribution of "Skill1&2" values given each possible combination of the values of its parents, "Skill1" and "Skill2," as if it were true. The numerical values of all the conditional probability relationships for the examples in this presentation were approximated with results from Tatsuoka's (1983) "rule space" analysis of the data, using only students classified as Method B users.[4]

Figure 10 depicts base rate probabilities of skill possession and item percents-correct, or the state of knowledge one would have about a student known to use Method B, before observing any item responses. Suppose we observe a pattern of responses that has mostly correct answers to items that don't require Skill 2, but incorrect answers to most of those that do. This is a body of disparate evidence: right and wrong answers to items involving different skills in different combinations. Its evidential value is discerned through the relationships whose structure is depicted in the DAG and whose strengths and directions are expressed in the accompanying conditional probability distributions. (The network could be extended to accommodate evidence from even more disparate sources, such as teachers' observations or explanations of solutions, if conditional probabilities of their outcomes given potential values of the skill nodes could be assessed. The extended network might require variables to model the effects of important influences on the new observables, above and beyond the skill variables.) Assuming the veracity of this structure, Figure 11 shows how beliefs change after observing such a response pattern. In particular, the updated probabilities for the five skills required for various items under Method B 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.

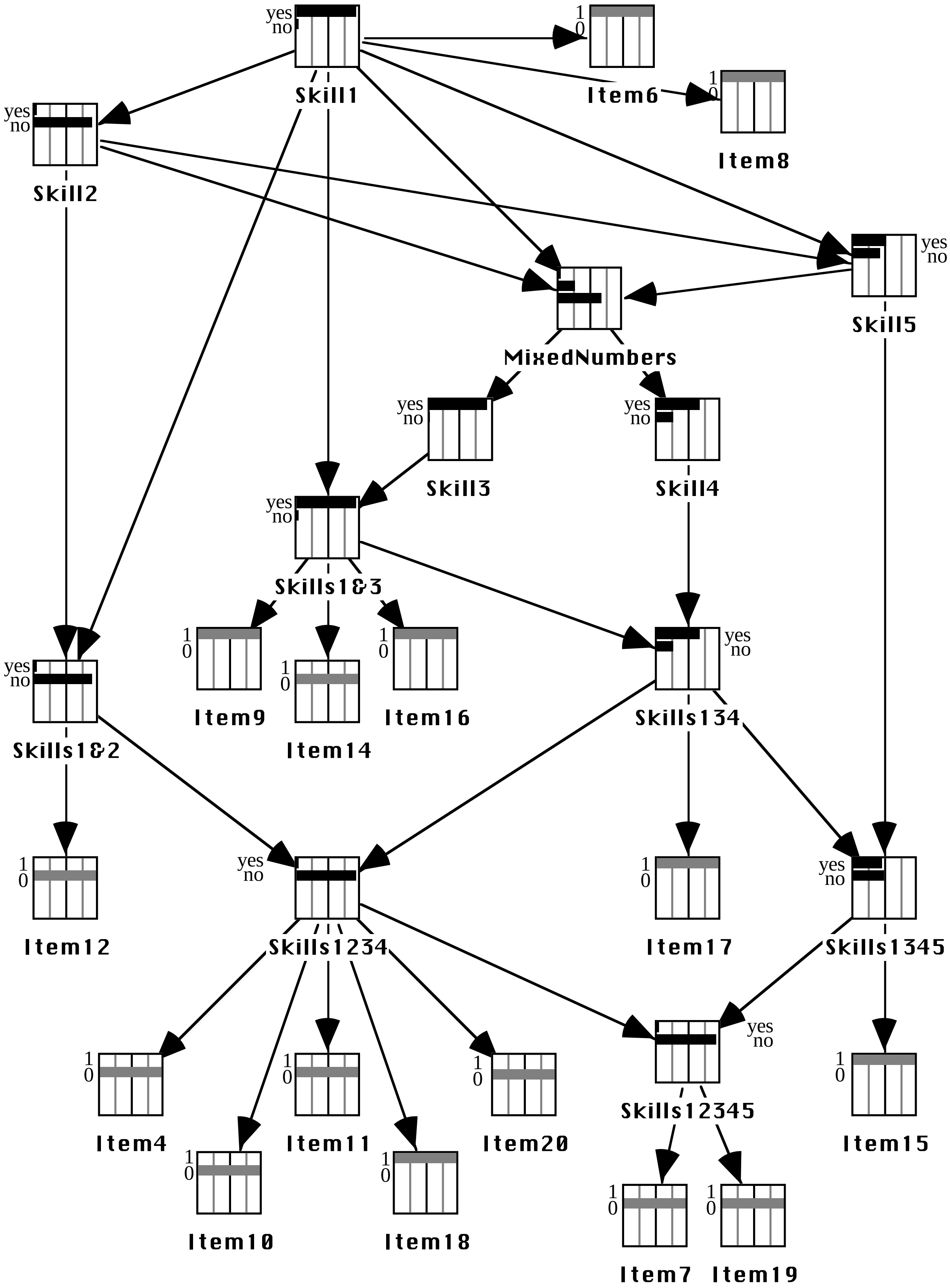
Figure 12 incorporates the Method B network and a similar network for Method A into a single network that is appropriate if we don't know which strategy a student uses. The evidential structure is a disjunction, not one of Wigmore's basic structures but as common in educational assessment as in everyday life: There are multiple routes to an outcome, and observing the

[4] Duanli Yan and I have also estimated conditional probabilities in this network with the EM algorithm and are currently working on Gibbs sampling characterizations of such networks.



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

Figure 10. Inference network for Method B, initial status.



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.

Figure 11. Inference network for Method B, after observing item responses.

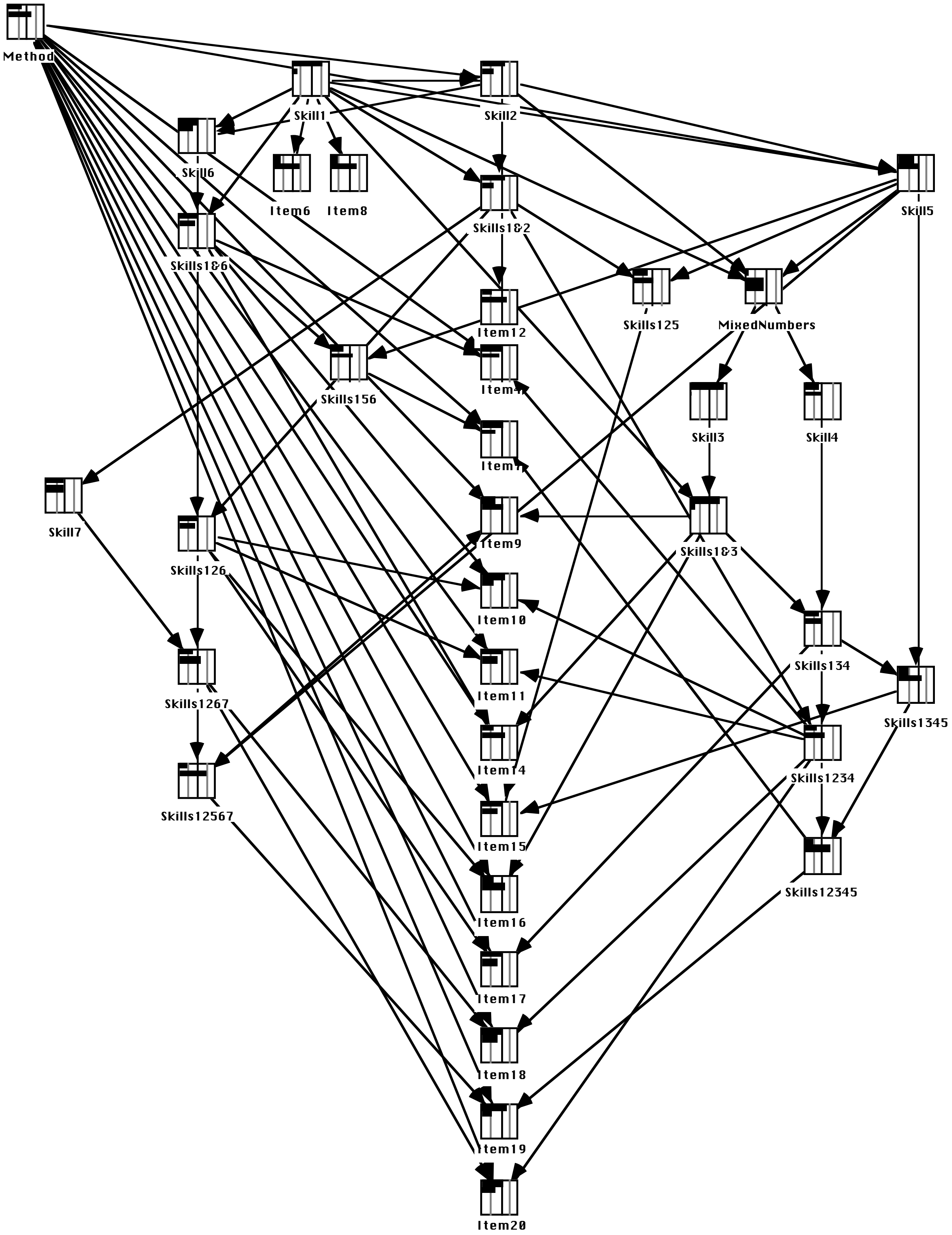


Figure 12. Directed acyclic graph for both methods.

> outcome alone does not indicate the route. Each item-response node now has three parents: minimally sufficient sets of subprocedures under Method A and under Method B, and the new node "Is the student using Method A or Method B?" By virtue of their demands, two items can have the same minimal sufficient set of skills under one method but different minimal sets under the other. Their responses are conditionally independent *only* given status on these minimally sufficient skill sets and the method with which they are attempted. We find that 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 the second kind wrong shifts belief toward Method B. The opposite pattern shifts belief toward the use of Method A. These are patterns in data that constitute noise, in the form of conflicting evidence, in an overall proficiency model, yet which constitute evidence, in the form of converging evidence, about strategy usage under the combined network—a conjecture that cannot even be framed within the overall proficiency 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. The student model could be extended by allowing for strategy switching (Kyllonen, Lohman, & Snow, 1984); that is, deciding whether to use Method A or Method B on an 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 various conditions, with "always use Method A" and "always use Method B" as extreme cases.

The Role of Conditionality

When the target inference is defined in terms of general behavioral tendencies over a specified domain of task situations, modeling responses as if conditionally independent given "average proficiency" as in Example 2 can be a useful expedient for characterizing the evidential value of observations. The evidence a task provides is posited to have the same character for all students, expressed through probabilities of potential responses x given θ. Obviously, however, any particular task might be relatively easy compared with other tasks for some students but relatively hard for other students, due, perhaps, to the different books they have read, courses they have taken, or experiences through which they have developed their proficiencies. Such interactions are a source of uncertainty with respect to inference about overall proficiency defined in this manner, and more extensive interactions further *degrade* the tasks' weight of evidence about overall proficiency. This is appropriately signaled in classical test theory by lower reliability coefficients and in IRT by lower slope parameters. From a constructivist perspective, these interactions are fully expected, since knowledge typically develops first in context, then is extended and decontextualized so that it can be applied across a broader range of contexts. This point of view can suggest a different student-model variable, a different target inference, and additional conceptual relationships to support that inference—a situation in which more extensive interactions can *enhance* the weight of evidence from task responses, to the extent that the differential patterns are expected outcomes of distinctions in a more variegated student model space.

> **Example 5: Assessing Proficiency in a Foreign Language**. The mileposts described in the American Council on the Teaching of Foreign Languages (ACTFL) *Proficiency Guidelines* for reading (American Council on the Teaching of Foreign Languages, 1989), excerpts of which appear in Table 5, are founded on empirical evidence and theories about the development of competence in acquiring information from text in a foreign language. Note the contrast between Intermediate readers' competence with texts "about which the reader has personal interest or knowledge" with Advanced readers' comprehension of "texts which treat unfamiliar topics and situation"—a distinction fundamental to the underlying conception of developing language proficiency. If we wish to assess students' proficiency in a foreign language, we encounter a fork in the road. Suppose, on one hand, the target of inference is overall proficiency with respect to a domain of tasks. We can predefine successful behavior on each task in the same way for all students regardless of their familiarity, administer a sample of tasks to a student, and thereby obtain direct evidence about expected behavior in the domain. Suppose, on the other hand, the target of inference is level of accomplishment with respect to the ACTFL Guidelines. If we know that the context of a given situation is familiar to one student but unfamiliar to a second, the same observed behavior from the two students holds radically different evidential import about their ACTFL levels. This example shows in a simple case how the machinery of probability-based inference can be applied when auxiliary information conditions the evidential value of students' performances.
>
> Contextual dependencies between situations and individuals can be incorporated into a Bayesian inference network by extending the structure beyond nodes that characterize the situation only from an "objective" point of

Table 5

Excerpts From the ACTFL Proficiency Guidelines for Reading

Level	Generic description
Novice-Low	Able occasionally to identify isolated words and/or major phrases when strongly supported by context.
:	:
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]
:	:
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. . . . [emphasis added]

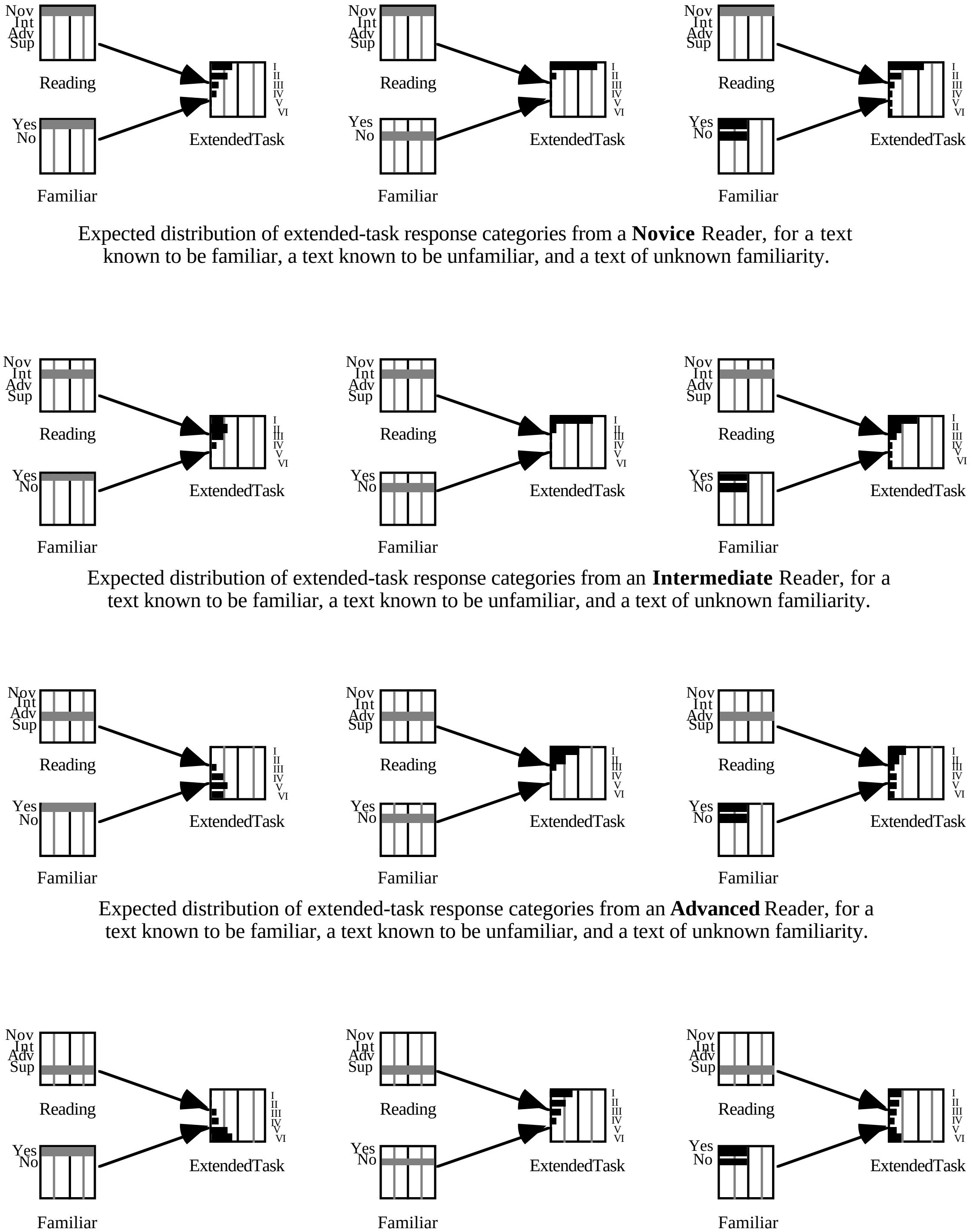
Note. Based on the *ACTFL Proficiency Guidelines,* American Council on the Teaching of Foreign Languages (1989).

view that pertains equally to all students. Nodes are introduced that vary across students in accordance with their points of view—for example, whether a student is familiar with the topic upon which a reading passage is based—and are modeled as additional parents of observable responses. Consider the following situation:

- The single student-model variable θ has four ACTFL levels, Novice, Intermediate, Advanced, and Superior.
- The observed variable x, a response to a passage based on a particular book, is rated in a five-category scale of quality, with levels denoted I, II, …, V.
- The student is characterized as either familiar or unfamiliar with the book in question, indicated by the auxiliary student/context familiarity variable y.

Figure 13 illustrates expectations about x as a function of given values of θ and y, or the flow of deductive reasoning. Note the different expectations when the student is and is not familiar with the context. Even students in the Superior category rarely perform well when the context is not familiar to them. When the student's level of familiarity is not known to an observer, the observer's expectations are a mixture of the two familiarity-known conditions and are consequently much more diffuse. (The mixture is weighted by the proportion of students in each category who are and are not familiar with the context; this illustration uses a 50-50 split.) Figure 14 shows the results of inductive reasoning from observing a low, medium, or high performance, under the conditions of (1) knowing the student is familiar with the context, (2) knowing the student is *not* familiar, and (3) *not knowing* whether the student is familiar. The task conveys much more evidence about reading competence when we know the student is familiar with the context, and very little when she is not. This kind of difference gains importance as tasks demand more time from students. The in-depth project that provides solid assessment information and a meaningful learning experience for the students whose prior knowledge structures it dovetails becomes an unconscionable waste of time for students for whom it has no connection.

If tasks provide so much more information when we know that the student is familiar with the context, why don't we always determine familiarity? The answer depends on the purpose of assessing and the cost of information to the assessor. Assessing a class of 30 fourth-grade students, a teacher can administer tasks related to what students have been studying and allow students to choose topics for projects. The teacher can generally arrange to observe data that can be interpreted under "familiarity=yes" conditions. A national testing program constrained to present the same tasks to 30,000 fourth-grade students generally cannot. Unlike a student's teacher, a distant observer lacks immediate and detailed information about contextual and situational student-by-task interactions.



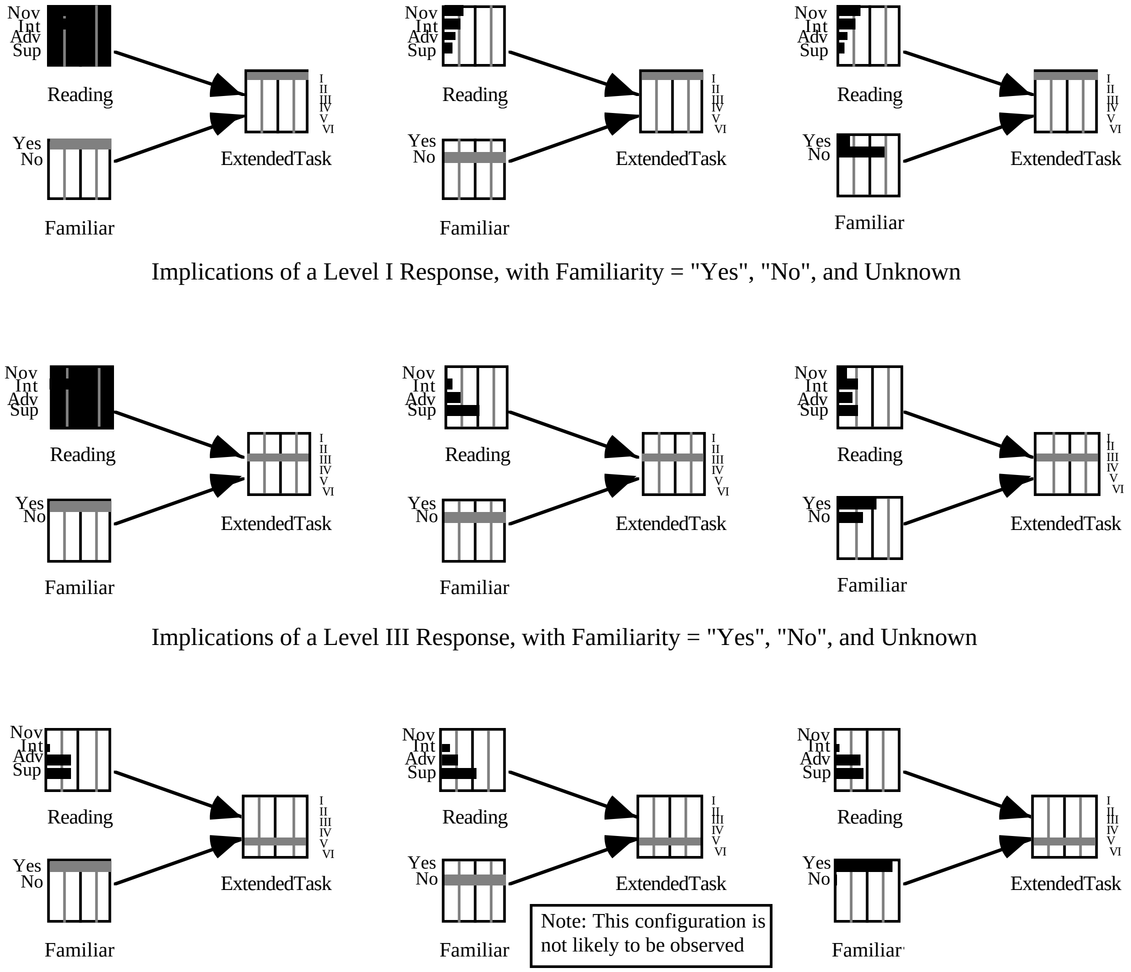
Expected distribution of extended-task response categories from a **Novice** Reader, for a text known to be familiar, a text known to be unfamiliar, and a text of unknown familiarity.

Expected distribution of extended-task response categories from an **Intermediate** Reader, for a text known to be familiar, a text known to be unfamiliar, and a text of unknown familiarity.

Expected distribution of extended-task response categories from an **Advanced** Reader, for a text known to be familiar, a text known to be unfamiliar, and a text of unknown familiarity.

Expected distribution of extended-task response categories from a **Superior** Reader, for a text known to be familiar, a text known to be unfamiliar, and a text of unknown familiarity.

Figure 13. Probabilities of task response categories, conditional on student competence and familiarity with text.



Implications of a Level I Response, with Familiarity = "Yes", "No", and Unknown

Implications of a Level III Response, with Familiarity = "Yes", "No", and Unknown

Implications of a Level V Response, with Familiarity = "Yes", "No", and Unknown

Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.

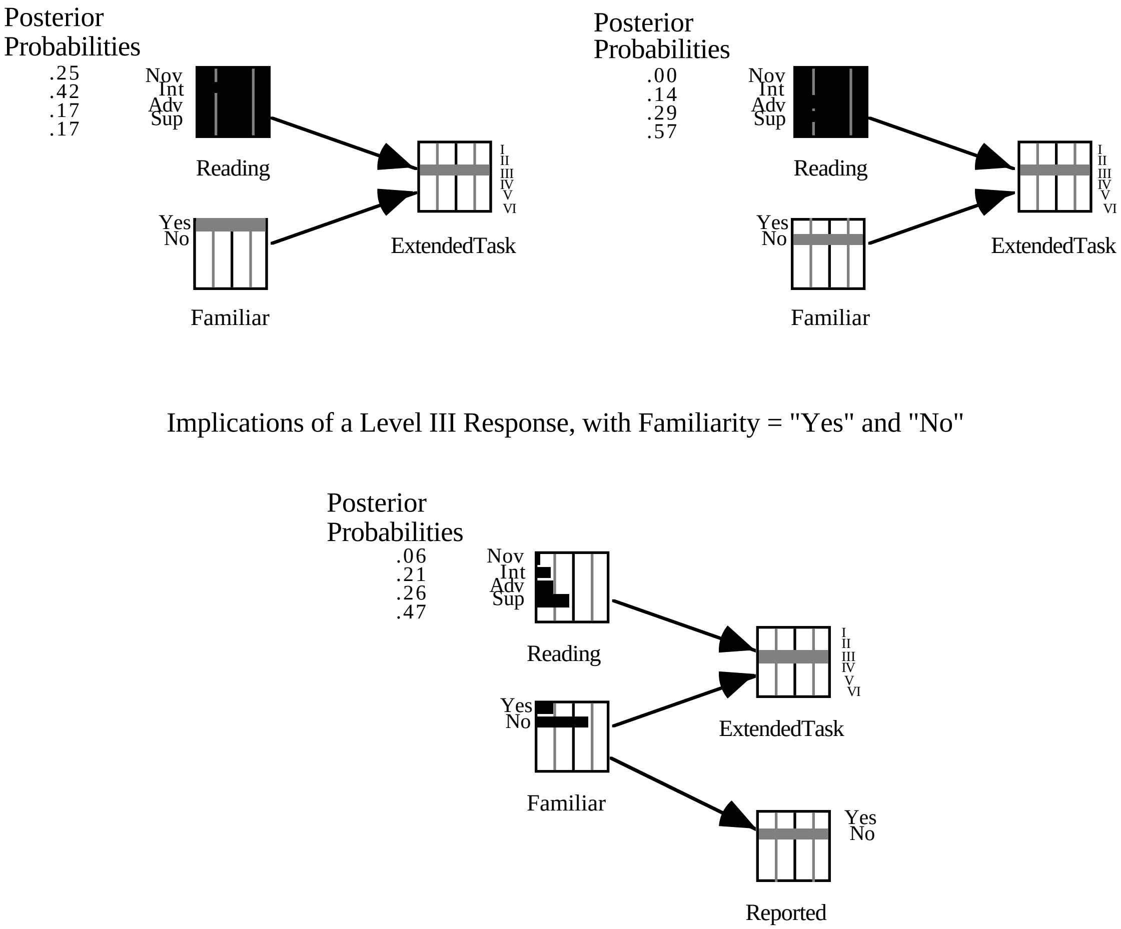
Figure 14. Posterior probabilities for student proficiency after observing task response, under various states of knowledge about task familiarity.

Some large-scale surveys gather "opportunity to learn" (OTL) information from teachers or students themselves in an attempt to shift inference from the default "familiarity=unknown" condition to either the "=yes" or "=no" condition (Platt, 1975). The good news is that OTL improves estimates of population-level relationships among schooling variables and attainment. The bad news is that OTL measures are not sufficiently dependable to be treated as "known with certainty" for individual students. Correlations between students' reports on background variables and independently verified values range from very low (~.2) to very high (~.9) (Koretz, 1992).

Figure 15 illustrates some consequences of uncertainty about auxiliary variables. Suppose that we did not ascertain familiarity directly, but obtained only a student's report. Suppose further that students who were truly unfamiliar with a context always reported they were unfamiliar, but 15-percent of the students who were truly familiar reported they were unfamiliar. The top two DAGs in Figure 15 repeat the inferences that follow if we know a student *is* familiar or *is not* familiar with the context. If a student is truly familiar, incorrectly reports he is unfamiliar, and we accept the report as a certain truth, then we mistakenly reason as shown in the top right DAG rather than the appropriate top left one. We would substantially overestimate his proficiency. The lower DAG adds a new node for the report. Its parent is true familiarity, and the conditional probability distribution when "familiarity=yes" is .85 for "report=yes" and .15 for "report=no." Conditioning on what we actually observe ("Report=yes" and "Task=III") accounts for this degree of uncertainty about true familiarity, and moderates the influence of the familiarity to a 87/13 mixture of "=no" and "=yes" familiarity-known conditions.[5] The result is an attenuated belief about proficiency that correctly reflects the average proficiency distribution among students with scores of III who report they are unfamiliar with the context. However, this distribution tends to understate slightly the proficiencies of those who are truly unfamiliar and still overstates substantially the proficiency of students who are familiar but report they are not. Depending on unsubstantiated reports in this manner would invite abuse in high-stakes "test as contest" applications; a student would raise his score by always claiming unfamiliarity whether it were true or not, even if the possibility of incorrect reports were accounted for on the average.

Tradeoffs between the potential value of evidence and the difficulties in ascertaining its credibility arise similarly in jurisprudence. American rules of evidence strictly limit hearsay testimony, or witnesses' claims about what a third party said. If that person isn't present, we can't be sure he made the statement in question; even if he did, we can't examine his demeanor when he says it, or cross-examine his motives and meanings. Although hearsay testimony can provide important information, it is generally excluded because it can also provide misinformation, be it guileless or self-serving, with little means for jurors to assess its credibility. In contrast, Swedish courts

[5] The relevant probabilities, now interpreted as the likelihood function, are p(report=no | familiarity=no)=1.00 and *p*(report=no | familiarity=yes)=.15, a ratio of 87/13 favoring familiarity=no.



Implications of a Level III Response, with Familiarity = "Yes" and "No"

Implications of a Level III Response, with *Reported* Familiarity = "No"

Note: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty.

Figure 15. Implications of a Level III response, with reported familiarity = "No."

have far fewer exclusionary rules of evidence and generally do admit hearsay evidence.[6] The side entering hearsay must be prepared in turn support its

[6] The Swedish system is closer to Bentham's ideal of "free evidence" proceedings. "To find infallible rules for evidence, rules which insure a just decision is, from the nature of things, absolutely impossible; but the human mind is too apt to establish rules which only increase the probabilities of a bad decision. All the service that an impartial investigator of the truth can perform in this respect is, to put legislators and judges on their guard against such hasty rules" (Bentham, 1825, p. 180).

credibility, however, through evidence and argumentation in further layers of catenation, to counter the doubts and counter-explanations the opposition advances. One must weigh the probative value of hearsay testimony against its requirements for support before deciding to use it.

Abductive Reasoning and Mathematical Probability

> There are perfectly satisfactory answers to all your questions. . . . But I don't think you understand how little you would learn from them. . . . Your questions are much more revealing about yourself than my answers would be about me.
>
> Peploe, Wollen, & Antonioni, 1975

A Bayesian inference network builds around theory-driven, deductive-reasoning structures—likely values of data given states of ultimate interest—in order to support subsequent inductive reasoning from realized data to probabilities of states. Yet abductive reasoning, apparently missing from the loop, is vital in two ways. First, just as a detective's and prosecutor's abductive reasoning provides the framework for the jury's inductive reasoning, insightful use of substantive theory is essential to construct the network. Secondly, while the network is a tool for reasoning deductively and inductively *within* the posited structure, abduction is required again to reason *about* the structure—to criticize and improve the structure, in response to mismatches between modeled and realized patterns. In the framework of mathematical probability, statistical diagnostic tools can highlight such anomalies as unexpected observations, departures from modeled conditional independences, and failures to capture salient features of data (Rubin, 1984). When we can model expected patterns with sufficient accuracy to be surprised when they don't occur, we open the door to learning; perhaps leading us to improve the way we collect data or to refine our statistical model, or, more profoundly, triggering a reconstruction of our conceptual model of the situation:

> 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. To the nature of that phenomenon, they usually provide no clues.
>
> Kuhn, 1970, p. 205

Inferring posterior distributions of parameters or predictive distributions of future observations within the framework of a model is analogous to a jury's guilty/not-guilty deliberation with respect to the prosecutor's story of the case. The establishment of a framework within which reasoning will take place facilitates communication, making explicit and public the structure of the argument and its grounding in evidence, and it secures credibility by separating the data-gathering and decision-making functions—but at the cost of narrowing the channel of what is communicated. Errors arise when the true state of affairs cannot be adequately approximated within the proffered framework. It is important to remember that the numerical probabilities that result from the use of Bayes theorem (and all the more when embedded in a complex network) depend on the posited structure. Only possibilities built into the model can end up with positive probabilities! Apparently precise numerical statements of belief prove misleading or downright embarrassing when it is later determined that the true state of affairs could not even be approximated in the analytic model.[7]

Two strategies from the mathematical-probability toolkit help address this problem in practice in educational assessment. One approach is to augment theoretically-expected unobservable states with one or more "catch-all" states to which increase in probability when unexpected patterns arise in observable data. Yamamoto's (1987) HYBRID model for item response data includes not only latent classes (such as those described in Example 3 above for proportional reasoning) that are associated with distinctive response patterns, but a catch-all class (the "IRT class") that merely characterizes examinees in terms of their overall tendency to answer items correctly. When response patterns occur that are unlike any of the patterns associated with the latent classes, the posterior probability for the catch-all class dominates; in this way, the model can express

[7] The House Select Committee on Assassinations assigned a 95% probability to the proposition that four shots were fired in the John Kennedy assassination, based on a dictabelt recording of sounds believed to have been recorded from a microphone on a police motorcycle in Dealy Plaza at the time of the incident. The sound patterns constituting the evidence, assumed to be echo impulses of shots during the six critical seconds, did in fact provide a much better match to experimentally-produced patterns for four shots than any other number of shots. But rock drummer Steve Barber discovered, faintly recorded on the dictabelt in the same time interval, words known to be spoken by Sheriff Bill Decker more than a minute after the assassination (Posner, 1993)—an observation that obviated any relationship between the putative echo impulses and the actual number of shots. The lesson is that the utility of numerical probabilities calculated within a posited inferential structure depends on the structure's fidelity to the real-world situation in question.

the fact that evidence may not support membership in *any* of the classes suggested by the associated substantive theory.[8] A second approach is to calculate indices of model misfit (in IRT, for example, Levine & Drasgow, 1982). While carrying out inference within a given probabilistic structure to update beliefs, indices are calculated to indicate how usual or unusual the observed data are under that structure: If higher level parameters took their most likely values in accordance with the observed datum, how likely would this datum be? Surprising observations are flagged, for it is here that actual circumstances may differ most severely from modeled circumstances.

> **Example 1: AP Studio Art Portfolios, continued**. A project can stimulate the kind of constructed learning or creative problem-solving thinking we wish to promote, yet fail nevertheless as an assessment tool unless we can abstract from the performance the critical evidence for the targeted inferences. It is necessary to establish a common framework of meaning among students and readers—shared standards for recognizing what is valued in performance and how it maps into the evaluative structure (Wolf, Bixby, Glenn, & Gardner, 1991). To this end, Carol Myford and I (Myford & Mislevy, 1995) have been studying the AP portfolio rating process from what might be called a "naturalistic" perspective and a "statistical" perspective. These two components of the project concern, respectively, the Baconian reasoning readers employ to assign ratings to portfolio sections, and Pascalian reasoning analyzing patterns among those ratings in a mathematical-probability framework—a partitioning in some ways analogous to that between the detective's realm and the jury's.
>
> In the "naturalistic" component, we identified 18 portfolios in the 1992 reading with a section that had received highly discrepant ratings from two readers. Currently, such occurrences are identified and rectified by a final rating from the chief faculty consultant; our motivation for discussing work that evoked discrepant ratings will become clear below. We discussed each section with two experienced readers to gain insights into the judging process in general, and into the features that made rating these particular portfolios difficult. The Wigmore chart shown above as Figure 1 above is based on one of these conversations. It would help this particular student understand why his Section B submission received the rating it did, and it would help other students, teachers, and new readers understand the kinds of evidence,

[8] Dempster-Shafer belief theory (Shafer, 1976) extends Bayesian inference in a manner that can also withhold support from all or some possibilities without having to assign support to other possibilities.

inference, arguments, and standards that underlie ratings more generally. However, more than 50,000 individual ratings were produced in the reading, and it is simply impossible to hold such discussions, let alone produce Wigmore charts, for each of them. A summary result for each, in the form of a numerical rating, provides the data for the complementary statistical perspective.

In the "statistical" component of the project, we used Linacre's (1989) FACETS model, a main-effects model for the log-odds of adjacent rating categories, to analyze patterns in the ratings. While IRT was invented to model regularities in examinees' overt behavior in common contexts considered invariant over people, FACETS uses similar mathematical structures to model regularities in readers' application of common standards to possibly quite different forms of evidence in different contexts from different students. How the student whose concentration was "angularity in ceramics" would fare in a domain defined by all possible concentration topics *is not* an inference of interest; the consistency with which different readers would map her particular accomplishments in "angularity in ceramics" into the common evaluative framework *is*. The data for each student were 13 scores on 0-to-4 scales, 3 from different readers on Section A (Quality), 2 from other readers on Section B (Concentration), and a total of 8 from each of two other readers on the four subsections of Section C (Breadth). The probability of a rating in category k on Scale h for a student with parameter θ from Reader j is modeled as

$$\mathrm{P}_{h,j,k}(\theta) = \frac{\exp\left[k\left(\theta - \xi_j + \eta_h\right) + \sum_{s=1}^{k} \tau_{sh}\right]}{1 + \sum_{t=1}^{K} \exp\left[t\left(\theta - \xi_j + \eta_h\right) + \sum_{s=1}^{t} \tau_{sh}\right]}. \tag{6}$$

The numerator is understood to be 1 for Rating Category 0; θ is a parameter for the portfolio, indicating a tendency over readers and sections to receive high or low ratings; ξ_j is the "harshness" parameter associated with Reader j; η_h is an "easiness" parameter for Section h; and τ_{kh}, for k=1,…,K, is a parameter indicating the relative probability of a rating in Category k as opposed to Category k-1 for the scale of Section h. Figure 16 graphs probabilities of response in each category of a 0-4 performance task as a function of θ. Figure 17 is a simplified version of the DAG for inference under this model.

The posterior distribution of the portfolio parameter, θ, summarizes the weight and direction of evidence provided by the 15 elemental ratings. Main effects of readers as to harshness or leniency are taken into account

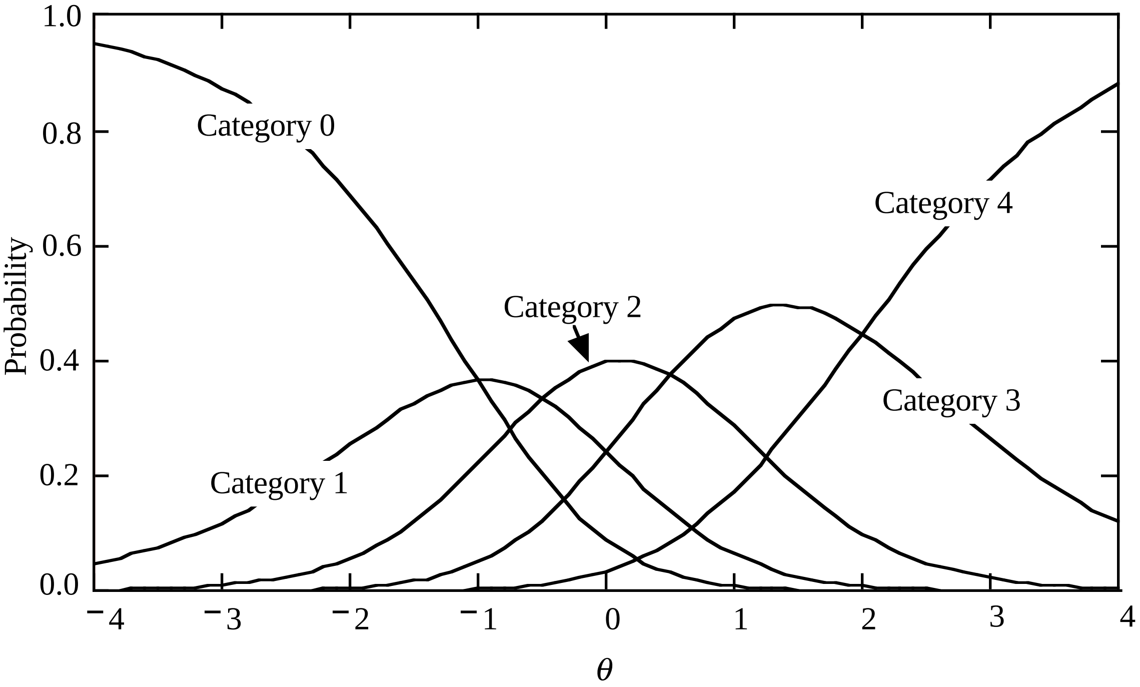


Figure 16. Probability of a response in categories as a function of θ, for a task with η=0, a reader with ξ=0, and τ=(1,.5,-.5,-2).

through ξ_j's, as are the average difficulties of the sections through η_h's. Figure 18 shows pairs of draws from the posterior distributions of the θ's of the 1992 portfolios, the spread away from the diagonal indicating the degree of uncertainty associated with the current configuration of readings. It is also possible to project through the model what the posterior precision of a portfolio parameter would be under different configurations of readings; say, one rating per section from different readers, two ratings for Sections A and B from the same two readers and two for Section C from two different readers, and so on (as in Cronbach, Gleser, Nanda, & Rajaratnam, 1972). This "pre-posterior" analysis is a tool for allocating a scarce resource (the expert readers' time) efficiently, as is done in adaptive testing with IRT.

While systematic reader effects can be taken into account, readers-by-portfolio interactions cannot be when, as in AP Studio Art, a reader rates a section only once; they therefore contribute uncertainty to the composite score. To what degree are these interactions caused by fatigue, by ambiguous directions to students or readers, by strongly idiosyncratic points of view, or different ways of integrating disparate aspects of accomplishment in the works within portfolio sections? Patterns of variation can be detected and quantified by statistical analyses, but the numbers cannot in and of

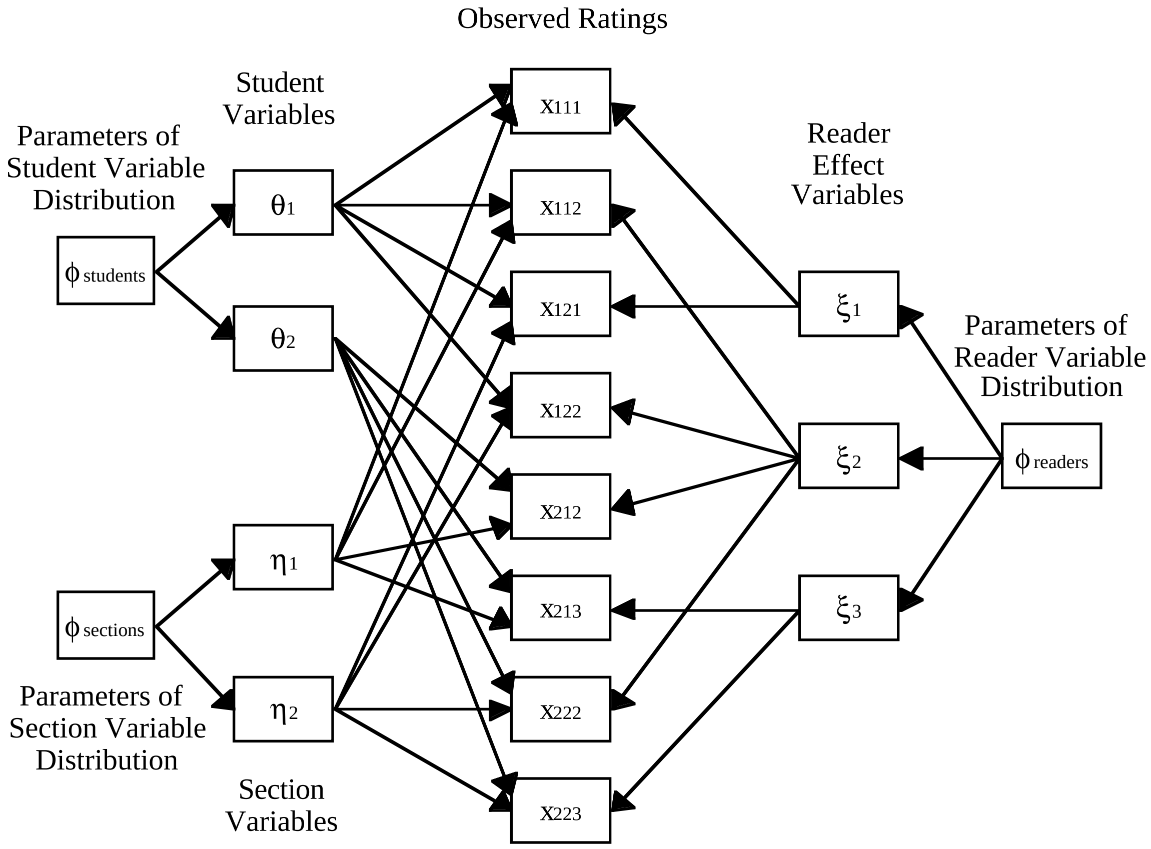


Figure 17. Directed acyclic graph for the AP Studio Art example.

themselves tell us how to improve reader training, sharpen the definition of standards, or distinguish aspects of accomplishment that should be rated separately. Since no one individual can become intimately familiar with all 50,000 rating processes, FACETS highlights particular reader/portfolio combinations that are especially unusual in light of the main effects, to help focus attention where it is most needed.

Statistical identification of outliers tells us where to look, but not what to look for. These cases are unusual precisely because the causes of variation we already understand do not explain them. Further insight requires information outside the statistical framework, to seek new hypotheses for previously unrecognized factors. When a discrepancy arises, how would Wigmore charts summarizing the abductive reasoning of two readers differ? Would one show themes the other missed, due perhaps to specialized knowledge about the glazes the student used? Or would similar themes appear, but with conflicting aspects integrated in accordance with differing priorities? Such analyses, as occurred informally in our discussions, can reveal opportunities to improve

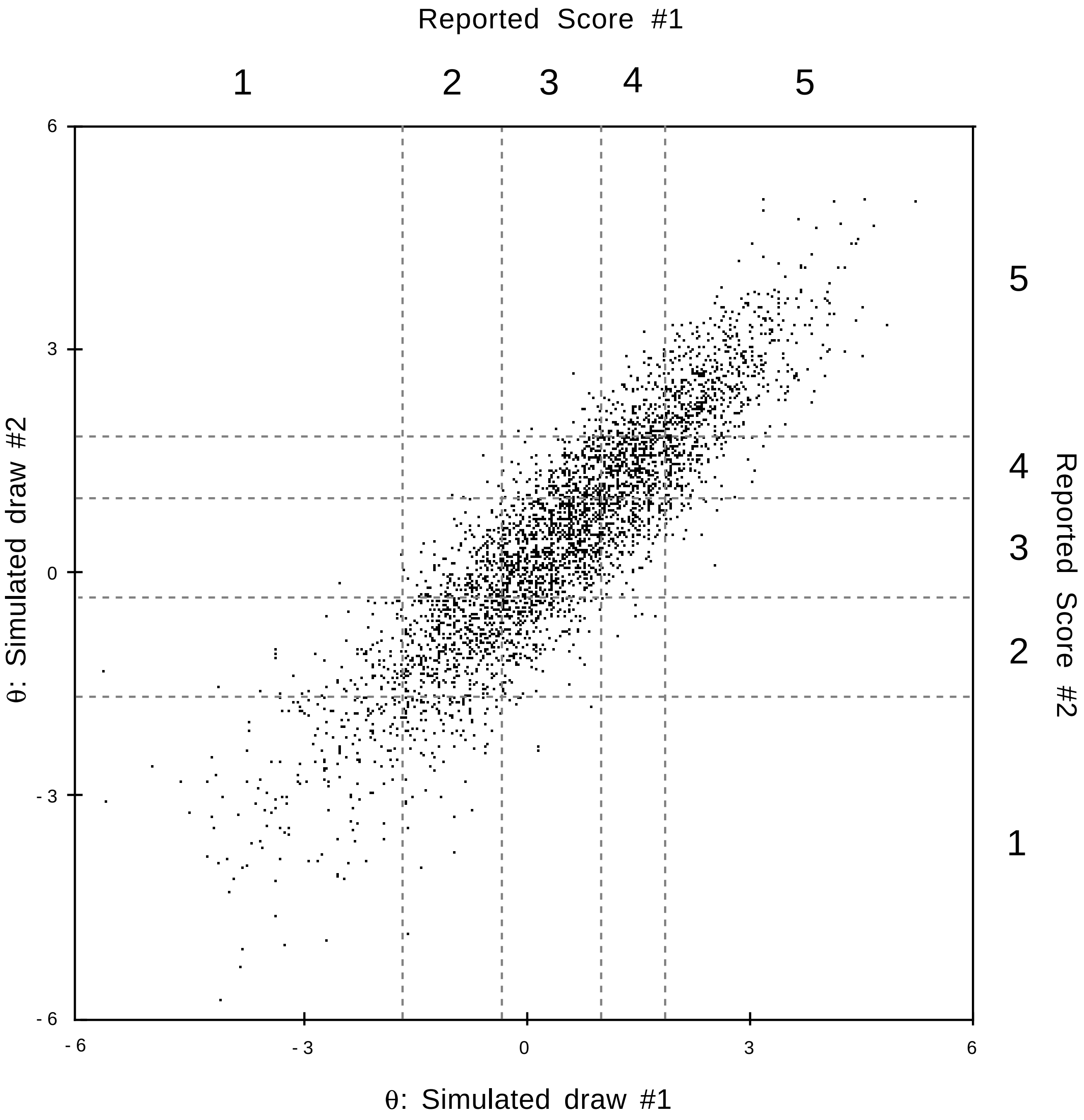


Figure 18. Two draws from the posterior distributions of portfolio parameters.

the evaluation system. Several avenues for possible exploration emerged in our project, including the development of verbal rubrics, particularly as a learning tool for new readers; having students write statements for the color and design sections, as for concentrations, to help readers understand the self-defined 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.

By working back and forth between statistical and naturalistic analyses, a common framework of meaning can be established, monitored, and refined over time. Readers' abductive reasoning from an open universe of possible student work leads to numerical ratings, through processes that can be made public through discussions, publications, or Wigmore charts concerning a range of representative examples. Once ratings have been obtained, statistical analysis can characterize evidence for inductive reasoning about typical cases within the system, and help identify atypical cases to trigger further abductive reasoning about the system itself. Mathematical tools originally developed under the mental measurement paradigm can thus be adapted to support inference in an assessment cast under a constructivist paradigm. By making public the materials and results of such a process, one can communicate the meaning and value of the work such assessments engender, and of the quality of the processes by which evidence about students' competence is inferred.

Conclusion

1. There is a close relation between the Science [of inference] and the Trial Rules [i.e., rules of evidence]—analogous to the relation between the scientific principles of nutrition and digestion and the rules of diet as empirically discovered and practiced by intelligent families.

2. The Trial Rules are, in a broad sense, founded upon the Science; but that the practical conditions of trials bring into play certain limiting considerations not found in the laboratory pursuit of the Science, and therefore the Rules do not and cannot always coincide with the principles of the Science.

3. That for this reason the principles of the Science, as a whole, cannot be expected to replace the Trial Rules; the Rules having their own right to exist independently.

4. But that, for the same reason, the principles of the Science may at certain points confirm the wisdom of the Trial Rules, and may at other points demonstrate the unwisdom of the Rules.

Wigmore, 1937, p. 925

Wigmore concluded that there are indeed general principles to guide and analyze evidentiary reasoning, but they alone are insufficient for the full range of issues of evidence and inference that arise in jurisprudence. To begin with, questions of what constitutes evidence cannot even be framed without conceptions of the nature of people and the nature of justice. Within a conceptual framework, determining whether and how to gather, admit, and evaluate data must weigh its evidential value against such considerations as the following: its tendencies to mislead jurors (e.g., hearsay testimony); costs of obtaining and supporting it (as this is written, genetic testing is potentially valuable, but often contentious and certainly expensive); and its feedback effects on the system (the Fifth Amendment protections against self-incrimination forgo highly relevant data, in order to discourage coerced confessions). Every general rule of evidence and every specific procedural decision must take such factors into account, but it should not, Wigmore argued, take them alone into account. Our chances of devising legal structures that strike appropriate balances among costs, rights, and correctness must surely increase as we more fully understand the implications of the tradeoffs we face. This includes, particularly and importantly, improving our understanding of the relationships between evidence and inference.

Educational assessment likewise takes place in social, political, theoretical, and personal contexts. Who collects and uses assessment data, for what purpose, at what costs, under what conception of competence, and with what feedback effects on curriculum and instruction? All of these issues impact assessment forms and practices—necessarily so, properly so. Yet assessment forms and practices, like rules of evidence, impact just as surely the weight and coverage of evidence that assessment data convey for the inferences and decisions they are meant to support. Apprehending the evidential value of assessment data requires (a) defining what we wish to accomplish, or our purposes for assessing; (b) specifying what we need to find out about students to achieve our purposes; and (c) constructing a principled framework in which we can evaluate and improve our efforts. As a general framework for reasoning in the presence of uncertainty, the paradigm of mathematical probability provides tools and concepts to further this end.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149,* 1-43.

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

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

Anderson, T. J., & Twining, W. L. (1991). *Analysis of evidence*. Boston: Little, Brown.

Andreassen, S., Jensen, F. V., & Olesen, K. G. (1990). *Medical expert systems based on causal probabilistic networks*. Aalborg, Denmark: Aalborg University, Institute of Electronic Systems.

Askin, W. (1985). *Evaluating the Advanced Placement portfolio in studio art*. Princeton, NJ: Educational Testing Service.

Bentham, J. (1825). *A treatise on judicial evidence*. London: Hunt & Clarke.

Bentham, J. (1827). *Rationale of judicial evidence*. London: Hunt & Clarke.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Cohen, L. J. (1977). *The probable and the provable*. Oxford: The Clarendon Press.

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.

de Finetti, B. (1974). *Theory of probability*. London: Wiley.

Diaconis, P., & Freedman, D. (1980). Finite exchangeable sequences. *The Annals of Probability, 8,* 745-764.

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

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: Lawrence Erlbaum Associates.

Greeno, J. G. (1989). A perspective on thinking. *American Psychologist, 44,* 134-141.

Gulliksen, H. (1961). Measurement of learning and mental abilities. *Psychometrika*, *26*, 93-107.

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: Lawrence Erlbaum Associates.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent trait variable models. *Annals of Statistics, 14*, 1523-1543.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kempf, W. (1983). Some theoretical concerns about applying latent trait models in educational testing. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 252-270). San Francisco: Josey-Bass.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York: Chelsea.

Koretz, D. (1992). *Evaluating and validating indicators of mathematics and science education* (RAND Note No. N-2900-NSF). Santa Monica, CA: RAND.

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

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

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrove (Eds.), *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press.

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.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Studies in social psychology in World War II, Volume 4: Measurement and prediction* (pp. 362-412). Princeton, NJ: Princeton University Press.

Levine, M., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

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.

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability of inference. *Annals of Statistics, 9*, 45-58.

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

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.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177-196.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Erlbaum.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, *54*, 661-679.

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.

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

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools*. New York: The Free Press.

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (ETS Research Rep. MS 94-05). 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.

Peploe, M., Wollen, P., & Antonioni, M. (1975). *The passenger*. New York: Random House.

Platt, W. J. (1975). Policy making and international studies in educational evaluation. In A. C. Purves & D. U. Levine (Eds.), *Educational policy and international assessment* (pp. 33-59). Berkeley, CA: McCutchen.

Posner, G. (1993). *Case closed: Lee Harvey Oswald and the assassination of JFK.* New York: Random House.

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

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12,* 1151-1172.

Schum, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance, 27,* 153-196.

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

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton, NJ: Princeton University Press.

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

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46*(2, Serial No. 189).

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-292.

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

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: Lawrence Erlbaum Associates.

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

Twining, W. L. (1985). *Theories of evidence: Bentham and Wigmore*. Stanford, CA: Stanford University Press.

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

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

Wigmore, J. H. (1937). *The science of judicial proof* (3rd ed.). Boston: Little, Brown.

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

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161-215.

Yamamoto, K. (1987). *A model that combines IRT and latent class models*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.