


# What is assessment really about, and how must it change?



Robert J. Mislevy  
Educational Testing Service

The Future of Education  
Northwestern University  
May 25-26, 2000

In learning a paradigm the scientist acquires theory, methods, and standards together, usually in an inextricable mixture.

Kuhn, 1970, p. 109

# Evidentiary Reasoning



It is amazing to me how many complex 'testing' simulation systems have been developed in the last decade, each without a scoring system.

The NBME has consistently found the challenges in the development of innovative testing methods to lie primarily in the scoring arena.

**Don Melnick, NBME**

We live in an age when we are still more adept at gathering, transmitting, storing, and retrieving information than we are at putting this information to use in drawing conclusions from it.

Kadane & Schum, 1996, p. xiv

The study of the principles of Evidence, for a lawyer, falls into two parts.

- One is Proof in the general sense, the part concerned with the ratiocinative process of contentious persuasion...
- 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...

Here we have been wrong; and in two ways.

Wigmore, 1937, pp. 3-4.

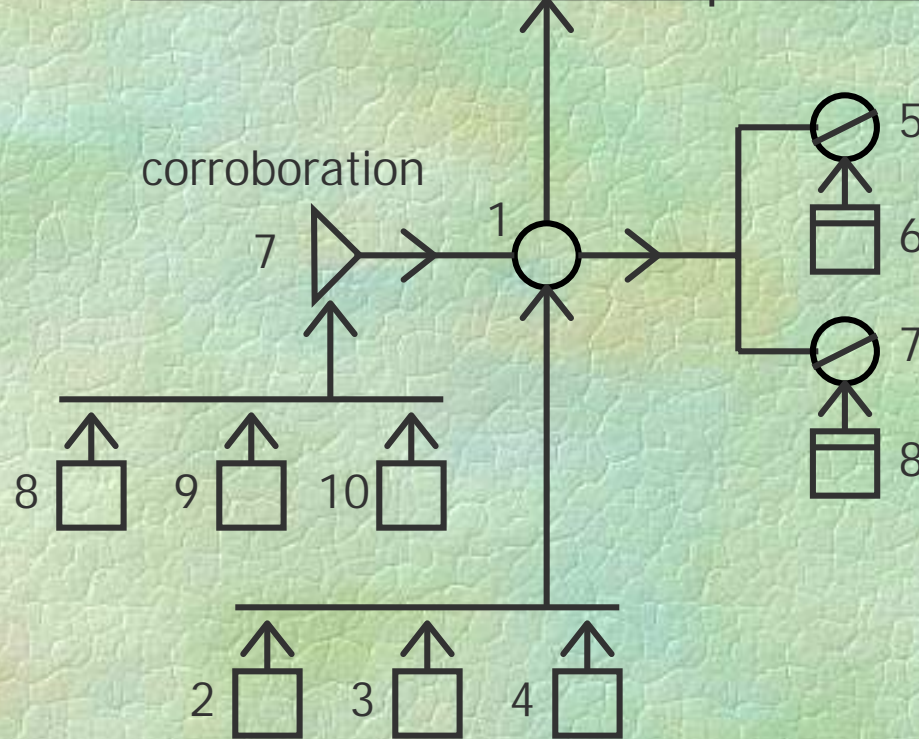
For one thing, there is, and there must be, a probative science—the principles of proof—independent of the artificial rules of procedure; hence, it can be and should be studied...

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.

# Issue: Did Y die of poison?

One line of  
Prosecution's  
argument



Defense's  
rebuttal

1. Y died, apparently in health, within three hours after the drink of whisky.
- 2-4. Y's wife and the Northington's witness to 1.
5. Y might have died from colic from which he had often suffered.
6. Witness testimony to Y's previous colic attacks
7. Colic would not have produced leg cramps and teeth-clenching; strychnine would have.
8. Y's wife and the Northington's witness to cramps and teeth-clenching.

[Example based on Schum, 1994; adapted from Wigmore, 1937]

Evidence forming the basis for probabilistic conclusions has three major properties or credentials that must be established: relevance, credibility, and inferential force.

No evidence comes with these credentials already established.

The task of establishing them rests, in part, on arguments or chains of reasoning we construct from the evidence to hypotheses or probable conclusions being considered.

Kadane & Schum, 1996, p. xi

Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. ...

[W]hat is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators—inferences about score meaning or interpretation and about the implications for action that the interpretation entails.

Messick, 1989, p. 13

# Assessment Design



Particular forms of tests and assessments represent particular forms of discourse, that is, they produce particular ways of talking and communicating with others about the schooling and education process.

Harold Berlak, 1992, p. 186

A construct-centered approach would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society.

Next, what behaviors or [performances should reveal those constructs, and

what tasks or situations should elicit those behaviors?

Messick, 1992, p. 17

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

# Cognitive Psychology



Summary test scores, and factors based on them, have often been thought of as “signs” indicating the presence of underlying, latent traits. ... An alternative interpretation of test scores as samples of cognitive processes and contents, and of correlations as indicating the similarity or overlap of this sampling, is equally justifiable ...

Whatever their practical value as summaries, for selection, classification, certification, or program evaluation, the cognitive psychological view is that such interpretations no longer suffice as scientific explanations of aptitude and achievement constructs.

Snow & Lohman, 1989, p. 317

In brief, [experts] ...

- (a) provide coherent explanations based on underlying principles rather than descriptions of superficial features or single statements of fact,
- (b) generate a plan for solution that is guided by an adequate representation of the problem situation and possible procedures and outcomes,
- (c) implement solution strategies that reflect relevant goals and subgoals, and
- (d) monitor their actions and flexibly adjust their approach based on performance feedback.

Baxter, Elder, & Glaser, 1996, p. 133

The fundamental character, then, of achievement measurement is based upon the assessment of growing knowledge structures, and related cognitive processes and procedural skills that develop as a domain of proficiency is acquired.

These different levels signal advancing expertise or passable blockages in the course of learning.

Glaser, Lesgold, & Lajoie, 1987, p.77

How much can testing gain from modern cognitive psychology? ...

So long as testing is viewed as something that takes place in a few hours, out of the context of instruction, and for the purpose of predicting a vaguely stated criterion, then the gains to be made are minimal.

Hunt, 1986, p. 22

# Situative Psychology



*Knowing*, in [the situative] perspective, is both an attribute of groups that carry out cooperative activities and an attribute of individuals who participate in the communities of which they are members...

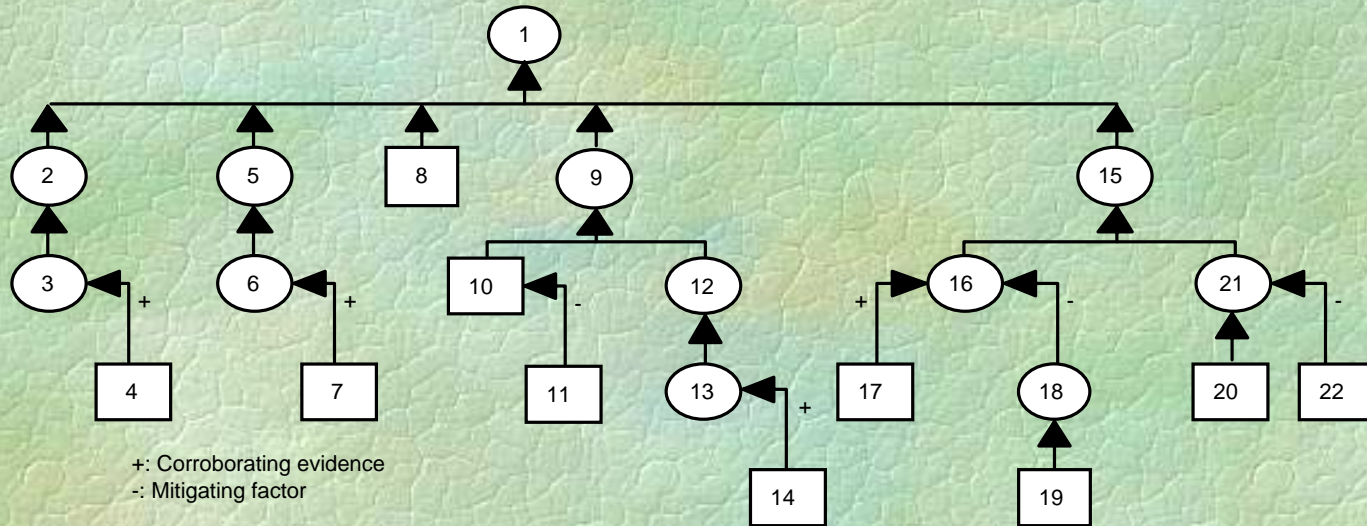
*Learning* by a group or individual involves becoming attuned to constraints and affordances of material and social systems with which they interact.

Greeno, Collins, & Resnick, 1997, p. 17

[A]ssessment is inevitably involved with questions of what is of value, rather than simple correctness. Questions of value require entry and discussion. In this light, assessment is not a matter for outside experts to design; rather, it is an episode in which students and teachers might learn, through reflection and debate, about the standards of good work and the rules of evidence.

Wolf et al, 1991, pp. 51-52

# Wigmore Chart for Rating an AP Studio Art Concentration



- 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 :
- 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.

[I]f one wishes to abstract a particular kind of activity from its cultural context and assert

(a) that it is a universal kind of achievement that differing peoples have mastered to a greater or lesser degree, and

(b) that one has a true theory of developmental stages in that domain,

then it is possible to do a kind of *conditional comparison* in which we can see how different cultures have organized experience to deal with that domain of activity.

Lab. of Comparative Human Cognition, 1982, p. 710

# Excerpts from the ACTFL Proficiency Guidelines for Reading

**Intermediate** Able to read consistently ... simple connected texts ... [that] impart basic information ... *to which the reader brings personal information and/or knowledge.*

**Advanced** Able to read ... prose ... with a clear underlying structure. ... *Comprehension derives not only from situational and subject matter knowledge but from increasing control of the language.*

**Superior** Able to read with almost complete comprehension ... expository prose on *unfamiliar subjects* and a variety of literary texts. Reading ability is not dependent on subject matter knowledge

# Probability-based reasoning



Probability is not really about numbers;  
it is about the structure of reasoning.

Glenn Shafer, quoted in Pearl, 1988, p. 77

A properly-structured statistical model embodies the salient qualitative patterns in the application at hand, and maps 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.

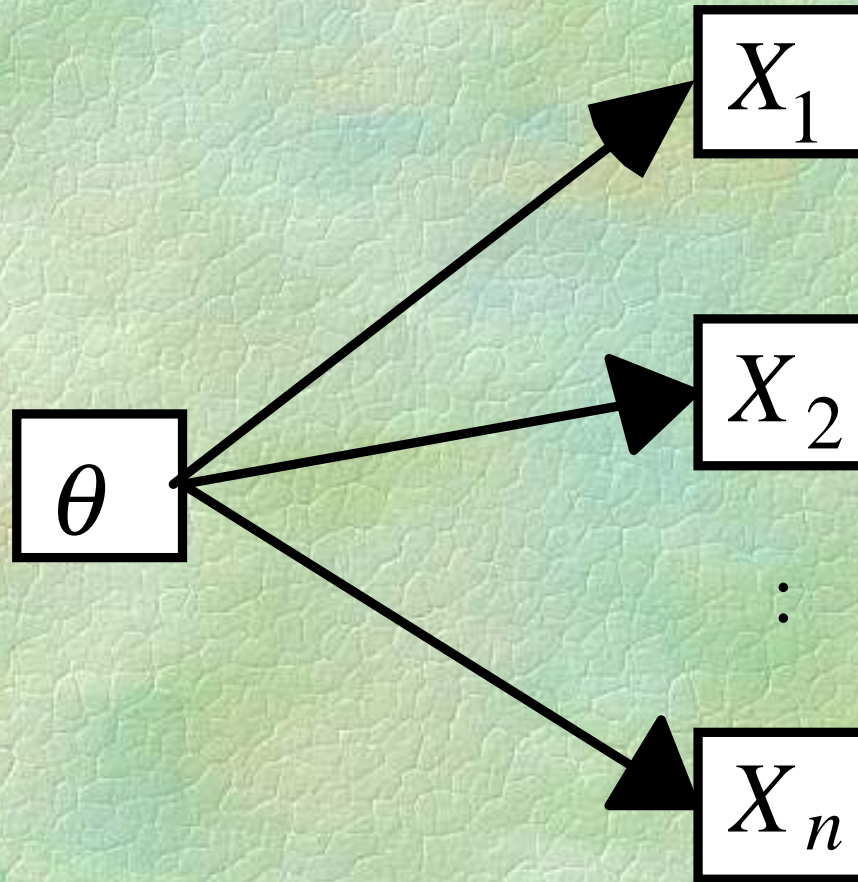
Mislevy, 1994

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.

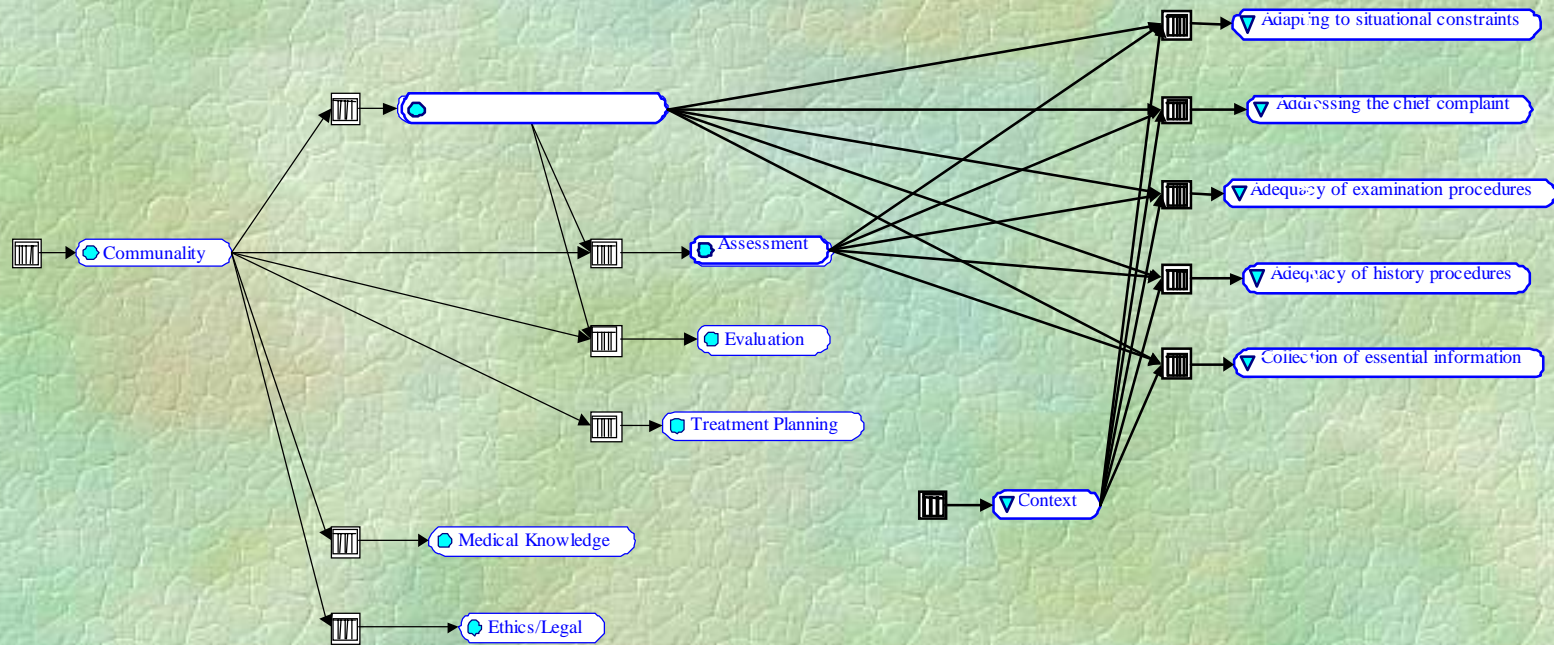
Pearl, 1988, p. 44

# Bayes net for the GRE



# DISC student model & evidence model

## Combined Bayes net



# Conclusion



In general, [experts] ...

- (a) provide coherent explanations based on underlying principles rather than descriptions of superficial features or single statements of fact.
- (b) generate a plan for solution that is guided by an adequate representation of the problem situation and possible procedures and outcomes...

Baxter, Elder, & Glaser, 1996, p. 133

# What are superficial features?

- Multiple-choice vs. performance assessments
- Modes of delivery
- Technology, in and of itself
- Reliability coefficients, in and of themselves

Different psychometric models might be employed [for performance assessments] to be sure, as well as different scoring procedures and rubrics, but such basic issues as validity, reliability, comparability, and fairness still need to be uniformly addressed.

This is so because *validity, reliability, comparability, and fairness are not just measurement issues, but social values* that have meaning and force outside of measurement wherever evaluative judgments and decisions are made.

Messick, 1992, p. 2

# What are the fundamentals?

- Transcendent principles of evidentiary reasoning,
- applied to inferences framed in terms of current and continually evolving educational psychology and educational practice,
- using current and continually evolving technologies to help gather and evaluate data.