Evidence-Centered Assessment Design

Robert J. Mislevy, Linda S. Steinberg,
and Russell G. Almond

Educational Testing Service
September 10, 1998

The work of the first author was supported by the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

Some scientific opportunities

Cognitive/educational psychology

- how people learn,
- organize knowledge,
- put knowledge to use.

Technology to...

- create, present, and vivify “tasks”;
- evoke, capture, parse, and store data;
- evaluate, report, and use results.

A Challenge

- How do you make sense of rich, complex data, for more ambitious inferences about students?

A Response

Design assessment from

generative principles ...

1. Psychology
2. Purpose
3. Evidentiary reasoning

    Conceptual design LEADS

    Statistics & technology FOLLOW

What is assessment?

Getting evidence about...

- what students know / can do / accomplish,
- from some theoretical perspective,
- under constraints,
- using some technologies,
- for some useful purpose.

Evidentiary Reasoning I: What inference is

- Inference is reasoning from what we know and what we observe to explanations, conclusions, or predictions.
- We always reason in the presence of uncertainty.

Evidentiary Reasoning II: ‘Data’ vs. ‘evidence’

- A *datum* becomes *evidence* in some analytic problem when its relevance to conjectures being considered is established.

- Conjectures, and the understanding of what constitutes evidence about them, emanate from the variables, concepts, and relationships of the domain.

Evidentiary Reasoning III: Reasoning from evidence

Evidence has three major properties that must be established:

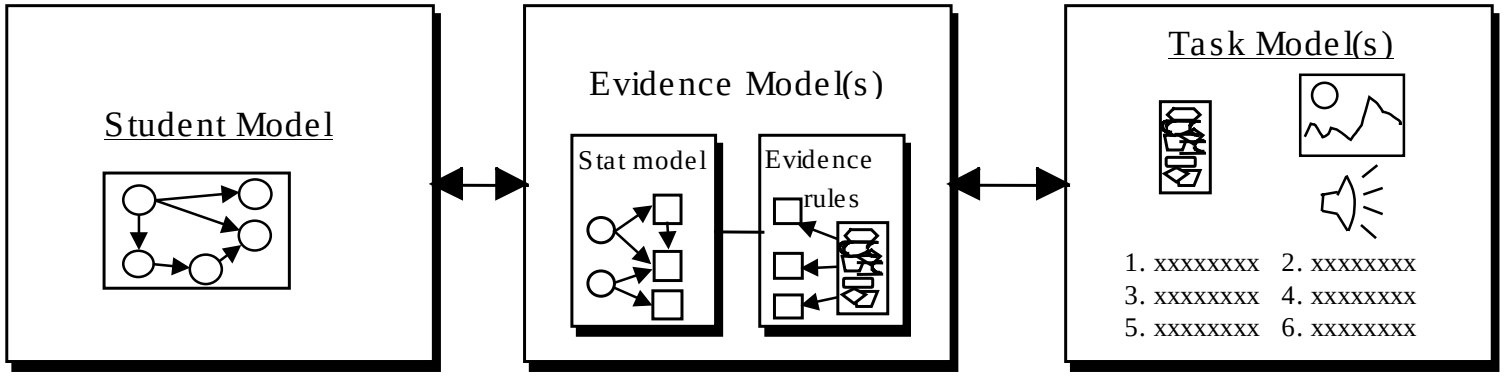
- relevance
- credibility
- inferential force (Kadane & Schum, 1996)

Some machinery for evidentiary reasoning

- Wigmore: “The science of judicial proof”
- Probability-based reasoning
- Bayesian inference networks

Principled Assessment Design

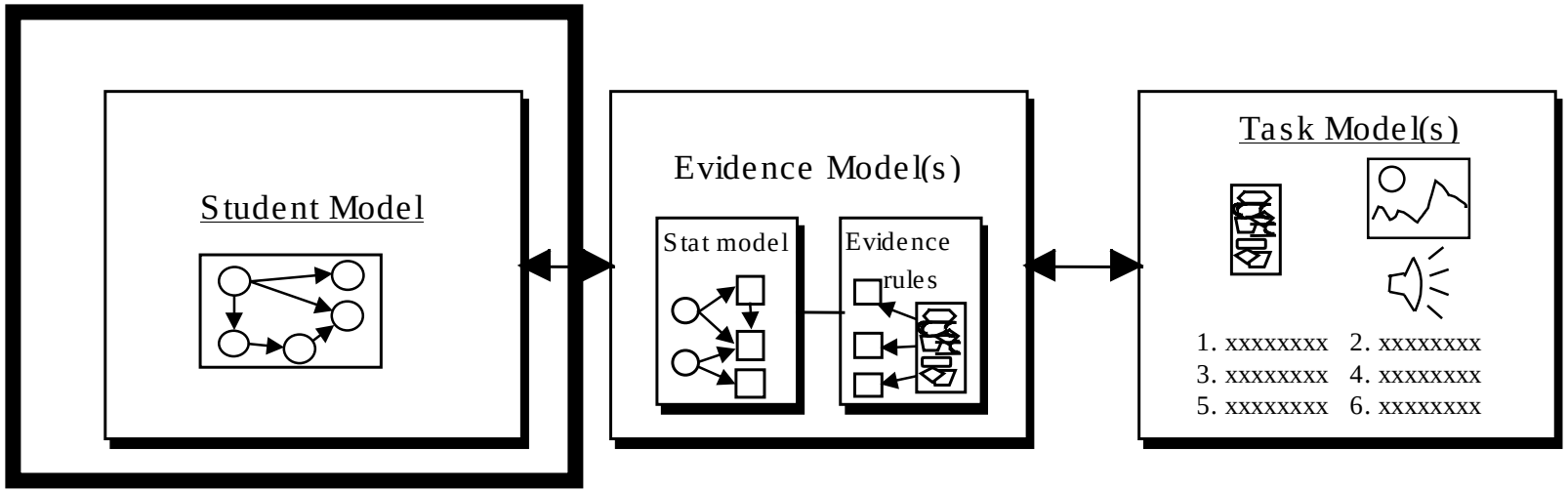
The three basic models



Evidence-centered assessment design

- What complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society?

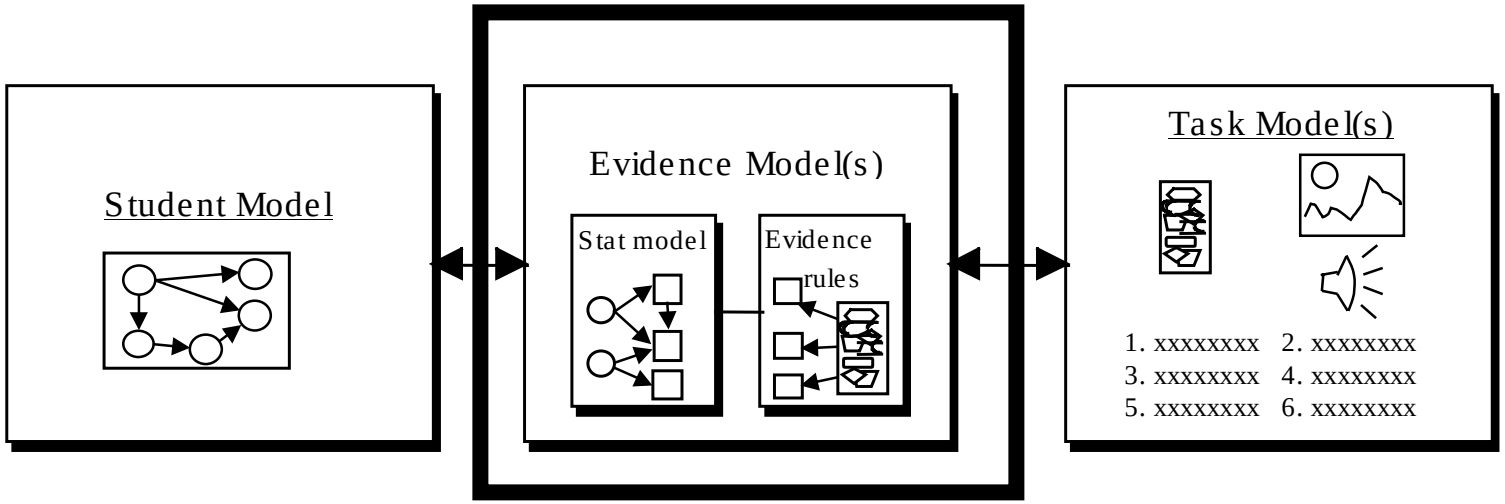
(Messick, 1992)



Evidence-centered assessment design

- What complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society?
- What behaviors or performances should reveal those constructs?

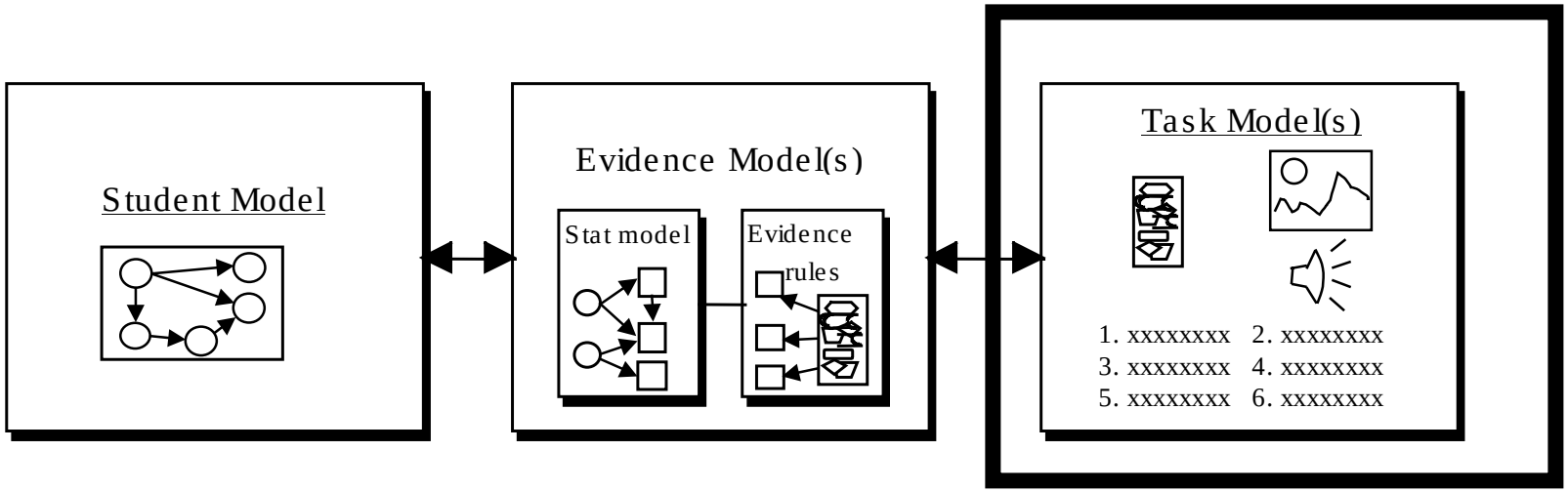
(Messick, 1992)



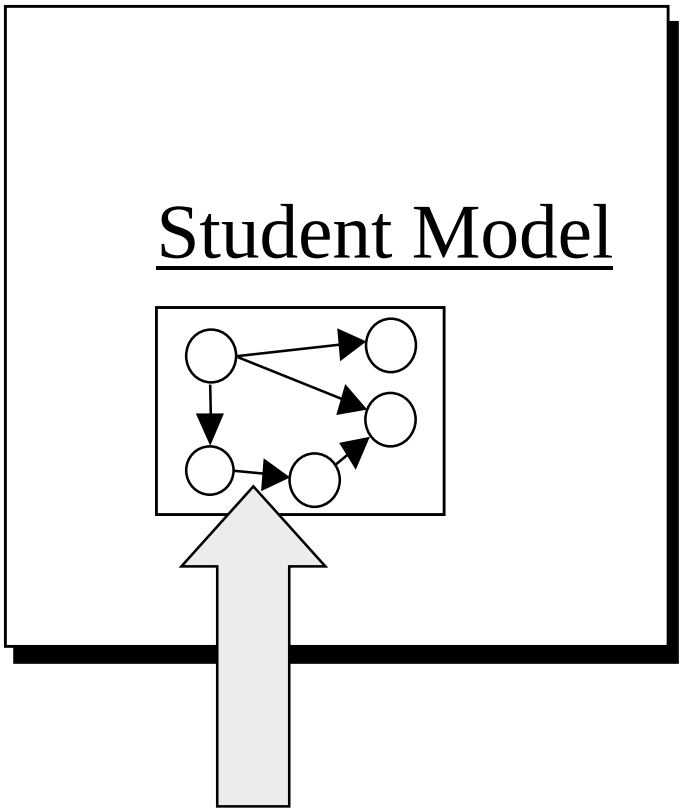
Evidence-centered assessment design

- What complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society?
- What behaviors or performances should reveal those constructs?
- What tasks or situations should elicit those behaviors?

(Messick, 1992)



The Student Model



Student-model variables describe characteristics of examinees

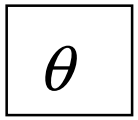
(knowledge, skills, abilities)

we want to make inferences about

(decisions, reports, diagnostic feedback, advice).

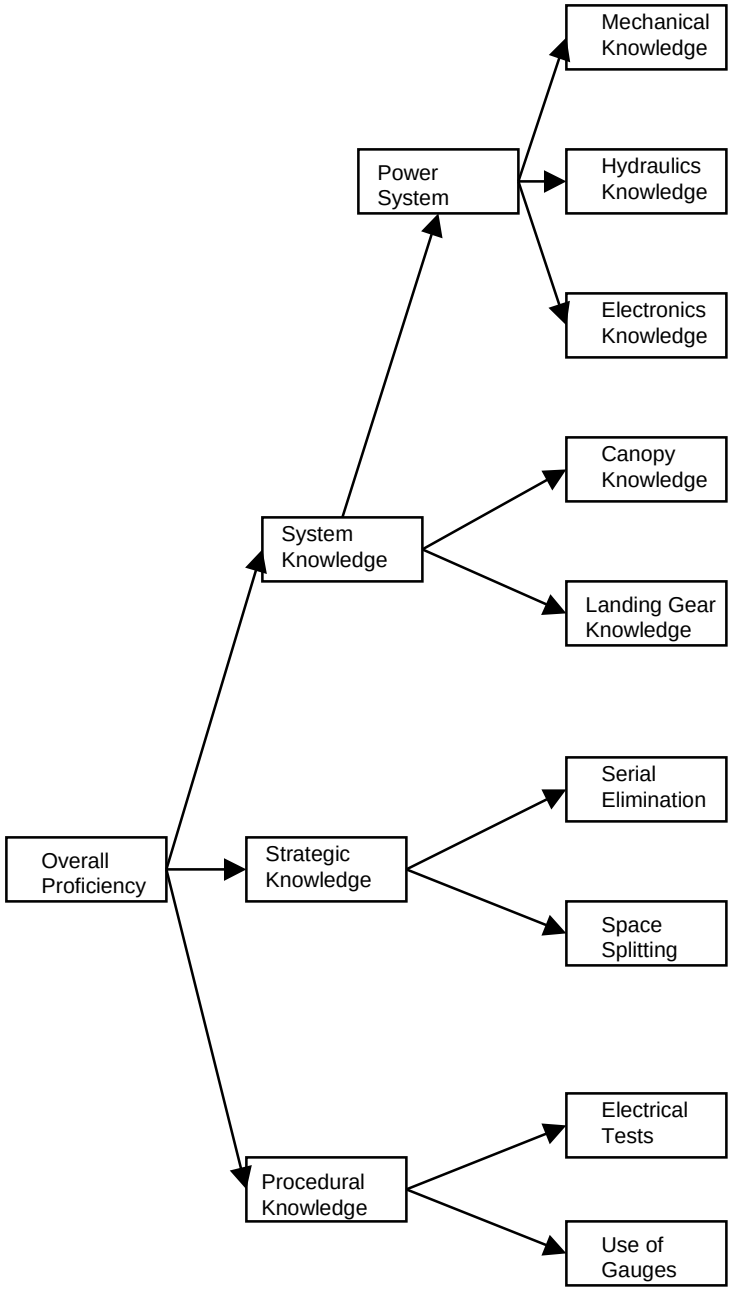
A fragment of a Bayes net.

Example a: GRE Verbal Reasoning



The student model is just the IRT ability parameter θ—

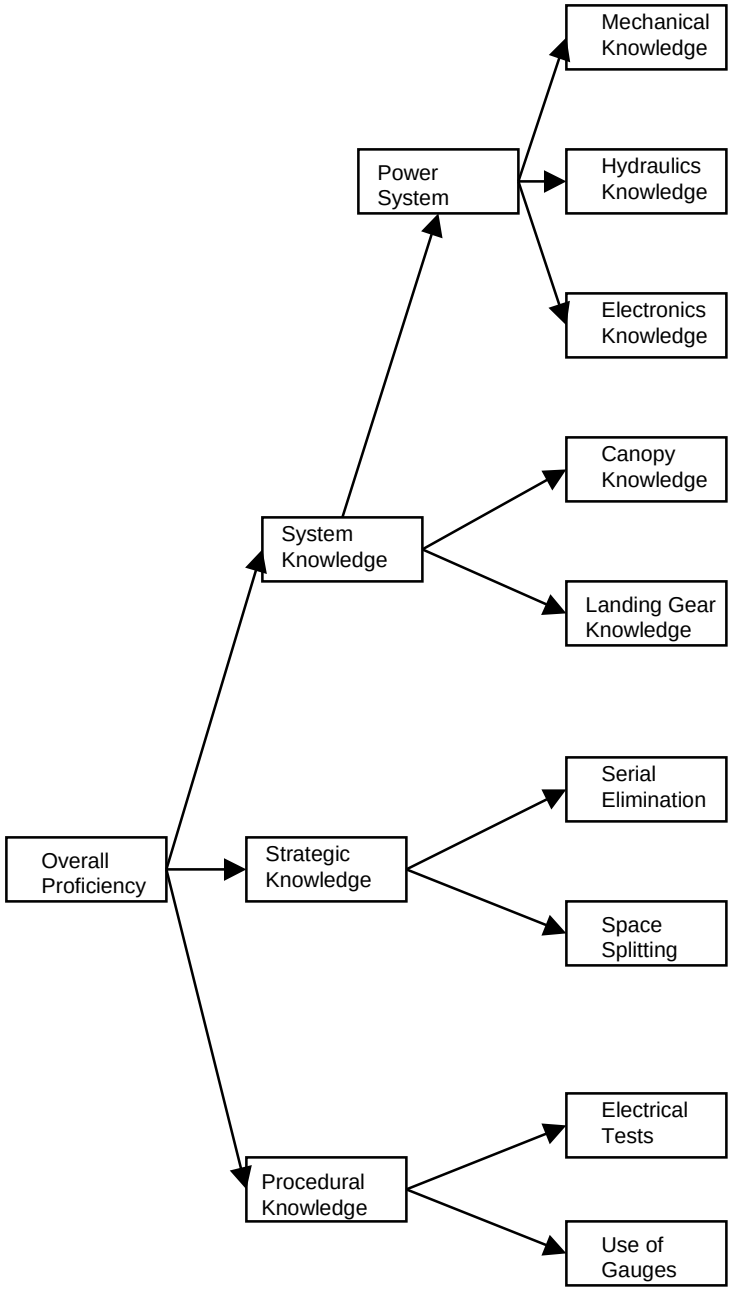
the tendency to make correct responses in the mix of items presented in a GRE-V.



Example b: HYDRIVE

Student-model variables in HYDRIVE

A Bayes net fragment.

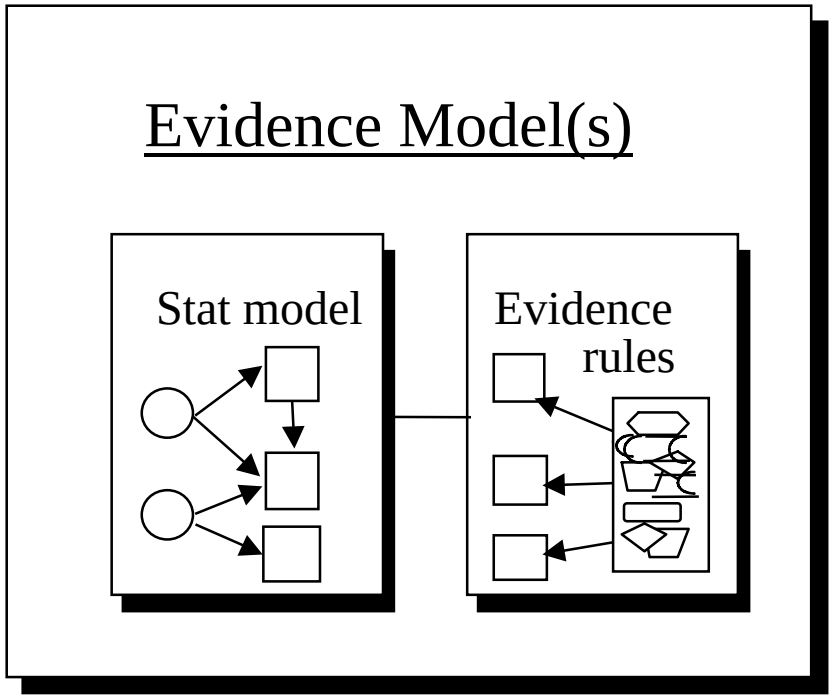


Example b: HYDRIVE

Student-model variables are derived from...

- Cognitive task analysis
- Instructional goals
- Instructional approach
- Simulator capabilities

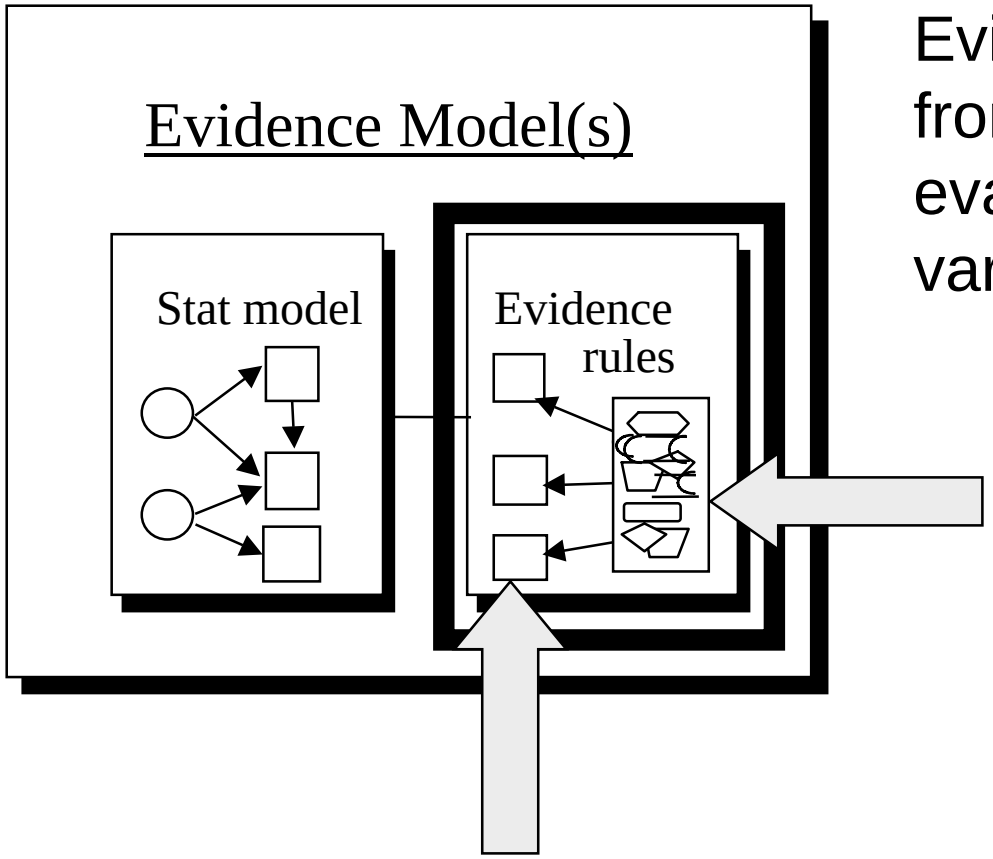
The Evidence Model(s)



Evidence-model variables concern features of student work.

An evidence model lays out the arguments for reasoning from what students say and do, to (1) what's important about it and (2) how it revises beliefs about the values of student model variables.

The Evidence Model(s)



Evidence rules extract features from a work product and evaluate values of observable variables.

Example a, continued: GRE-V

IF the area on the mark-sense answer sheet corresponding to the correct answer reflects more light by 10% than each area corresponding to the distractors, THEN the item response is correct.

Sample evidence rule

Example b, continued: HYDRIVE

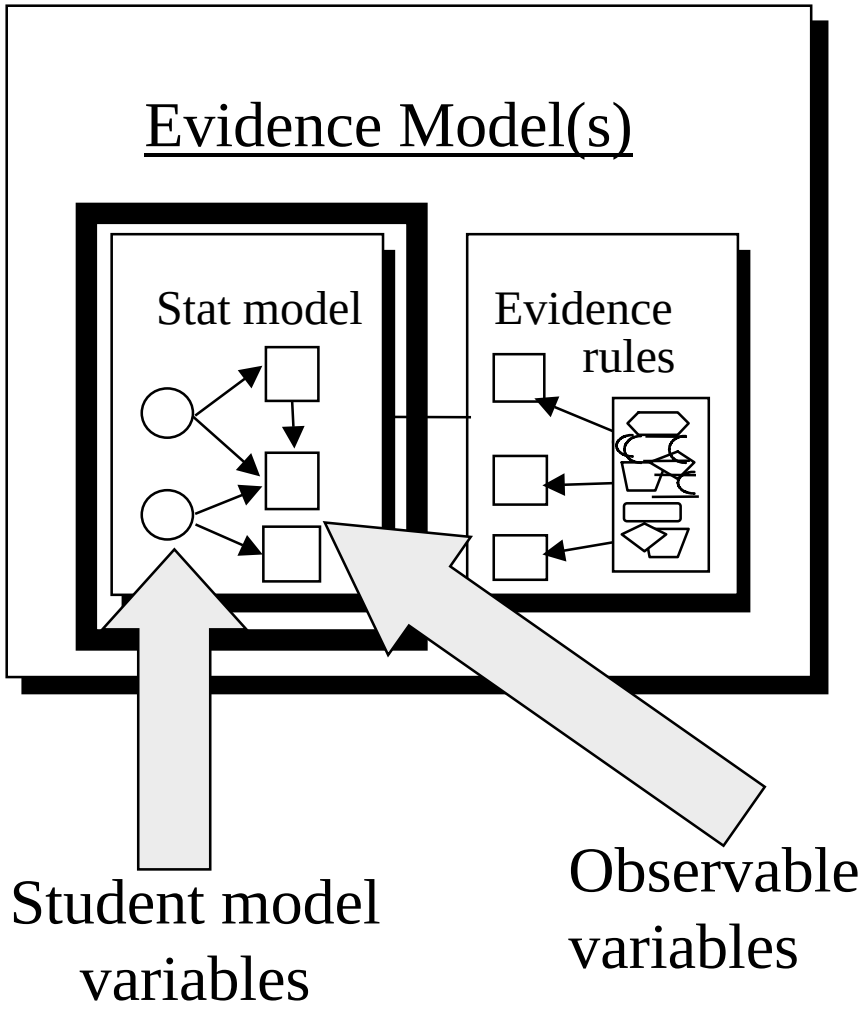
IF an active path which includes the failure has not been created and the student creates an active path which does not include the failure and the edges removed from the problem area are of one power class,

THEN the student strategy is splitting the power path

ELSE the student strategy is not splitting the power path.

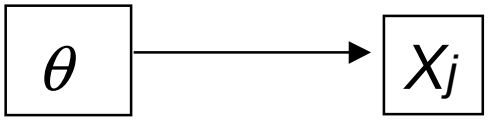
Sample evidence rule

The Evidence Model(s)



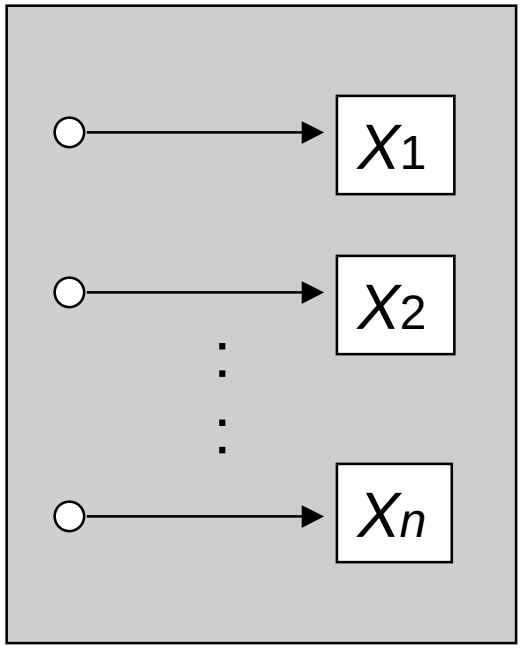
The statistical component expresses the how the observable variables depend, in probability, on student model variables.

Example a, continued: GRE-V



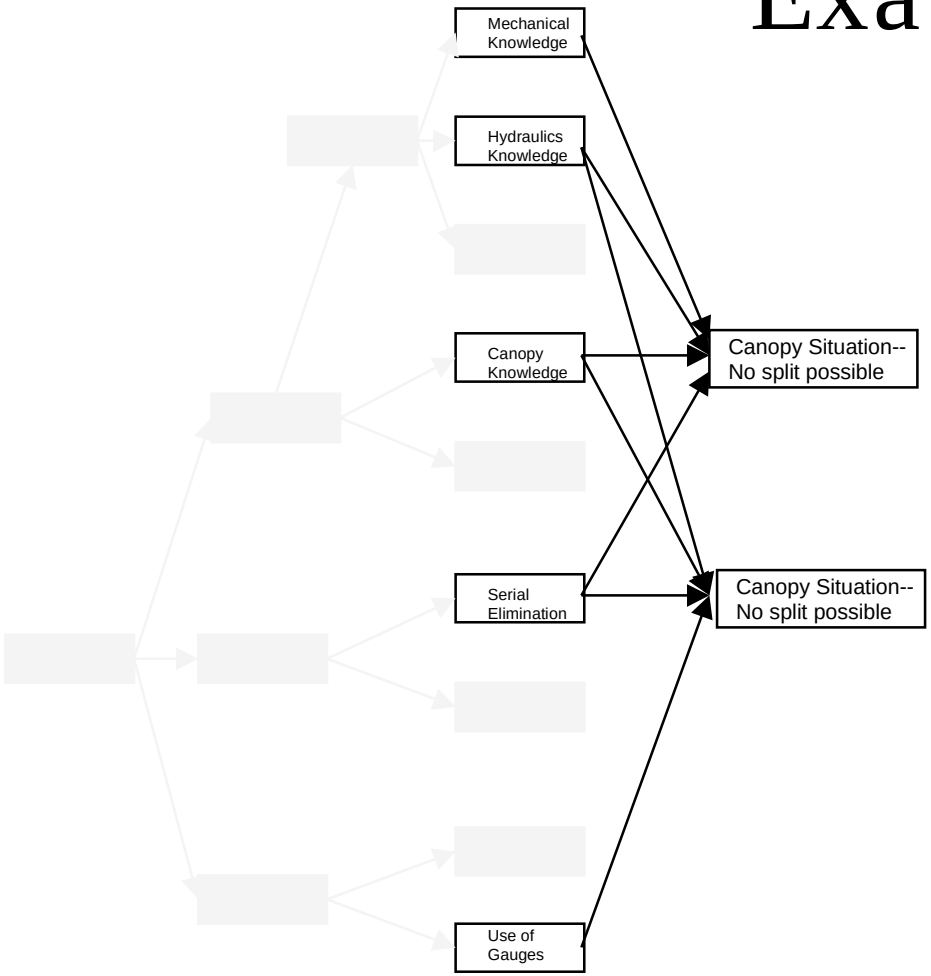
Sample Bayes net fragment

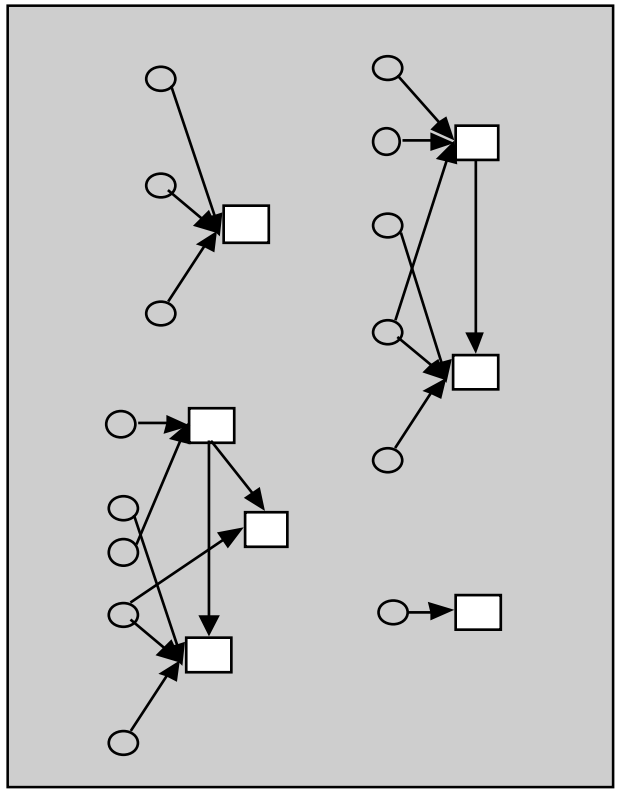
(IRT model & parameters for this item)



Library of fragments

Example b, continued: HYDRIVE

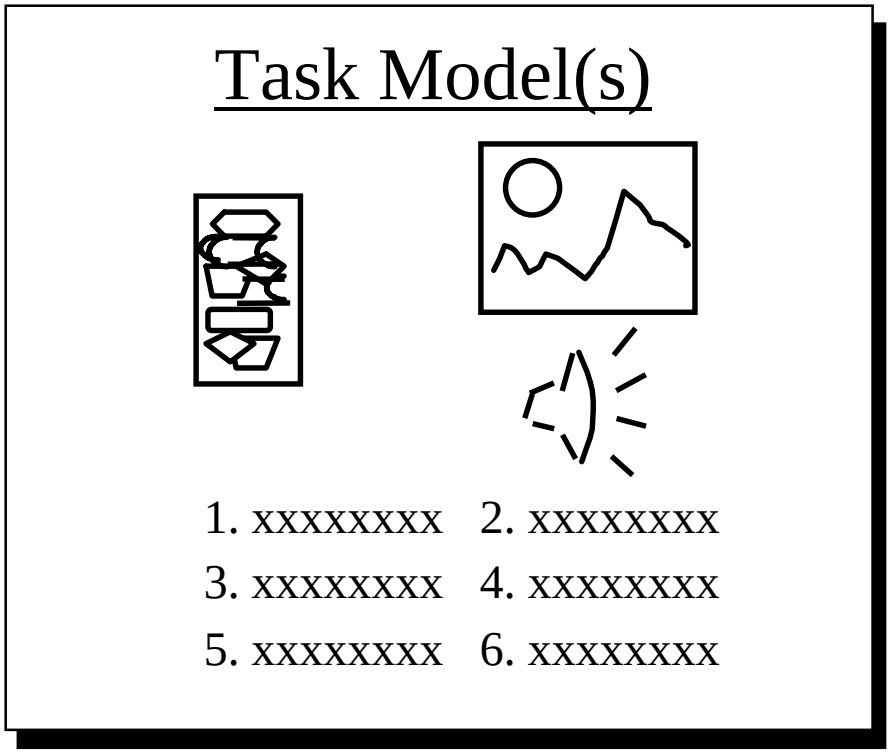




Sample Bayes net fragment

Library of fragments

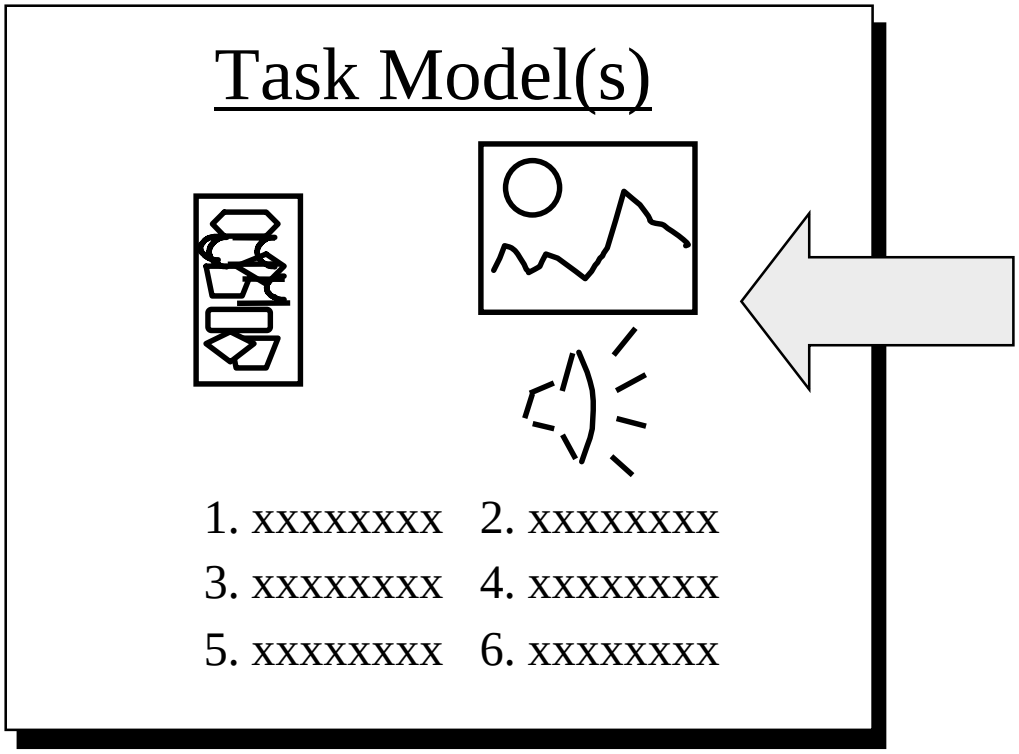
The Task Model(s)



Task-model variables concern features of tasks.

A task model provides a framework for describing and constructing the situations in which examinees act.

The Task Model(s)



Includes specifications for the stimulus material, conditions, and affordances--
the environment in which the student will say, do, or produce something.

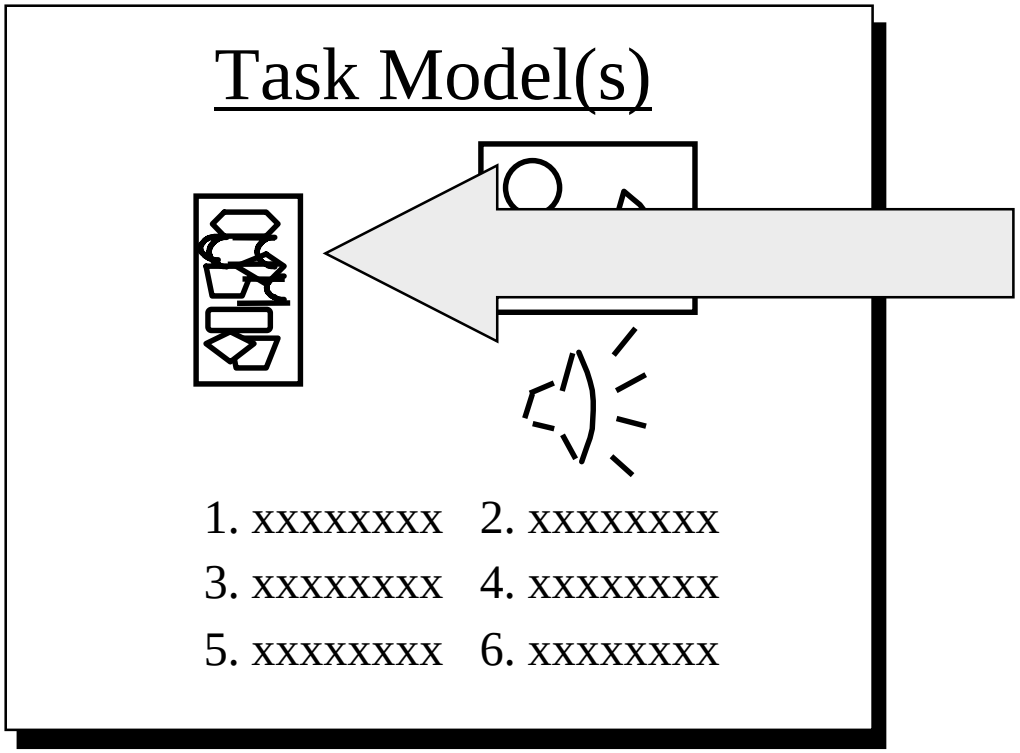
Example a, continued: GRE-V

- Content, format, cognitive-demand variables
- Variable that designates correct response
- Variables based on IRT model

Example b, continued : HYDRIVE

- Task = Selected fault in selected component. Some task models variables describe setup, initial state of system, stimulus materials, links to feedback & instruction.
- Simulator computes system state, provides outputs as function of aircraft state & student actions. Other task model variables describe aspects of changing state of components that will need to be computed and tracked.

The Task Model(s)



Includes specifications for the "work product": the form in which what the student says, does, or produces will be captured.

Example a, continued: GRE-V

- Work product is a pattern of filled-in answer bubbles.

Example b, continued : HYDRIVE

- Work product #1 is a file containing the sequence and time of actions taken by the student.
- Work product #2 is the state of the aircraft hydraulic simulator model after all of the actions (including part replacements, switch settings, etc) have been completed.

Conclusion

- There has been good progress in methods for gathering and using data in familiar forms of assessment.
- There are gaps between assessment users, policy makers, assessment innovators, & test theory specialists.

Conclusion

We can attack new assessment challenges by working from generative principles of assessment design:

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