How Technology and Psychology Challenge Educational Measurement

Robert J. Mislevy

Frederic M. Lord Chair in Measurement & Statistics
Educational Testing Service

Presented at the Conference on Measuring and Assessing Skills
October 1-2, 2015, The University of Chicago

Two Assessment Tasks

Pet Shop Display

Arturo is planning the parakeet display for his pet shop. He has five parakeets, Alice, Bob, Carla, Diwakar, and Etria. Each is a different color; not necessarily in the same order, they are white, speckled, green, blue, and yellow. Arturo has two cages. The top cage holds three birds, and the bottom cage holds two. The display must meet the following additional conditions:

Alice is in the bottom cage.
Bob is in the top cage and is not speckled.
Carla cannot be in the same cage as the blue parakeet.
Etria is green.
The green parakeet and the speckled parakeet are in the same cage.

1. If Carla is in the top cage, which of the following must be true?
   a) The green parakeet is in the bottom cage.
   b) The speckled parakeet is in the bottom cage.
   c) Diwakar is in the top cage.
   d) Diwakar is in the bottom cage.
   e) The blue parakeet is in the top cage.

Demo clip of SimCityEDU**: Pollution Challenge!**

By GlassLab (http://glasslabgames.org/)

The Standard Educational Measurement Paradigm (SEMP)

The Standard Educational Measurement Paradigm (SEMP).

- The goal is "**measuring a construct**" (θ)
- framed in trait or behaviorist psychology.
- Usually a single overall measure is desired.
- Each **task** (often an **item**) is a self-contained situation
- that evokes a **response** meant to provide evidence about the construct.
- Each response is evaluated to provide an **item score.**
- A **test score** accumulates evidence over items, usually summing item scores, sometimes through a model such as item response theory (IRT).

Sam Messick on Assessment Design

- [W]hat complex of knowledge, skills, or other attribute should be assessed...
- Next, what behaviors or performances should reveal those constructs, and
- what tasks or situations should elicit those behaviors?

Messick, 1994

Challenges and Opportunities re

Psychology

Snow & Lohman on "Measuring Traits"

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

An alternative interpretation of test scores as samples of cognitive processes and contents … is equally justifiable and could be theoretically more useful.

The evidence from cognitive psychology suggests that test performances are comprised of complex assemblies of component information-processing actions that are adapted to task requirements during performance.

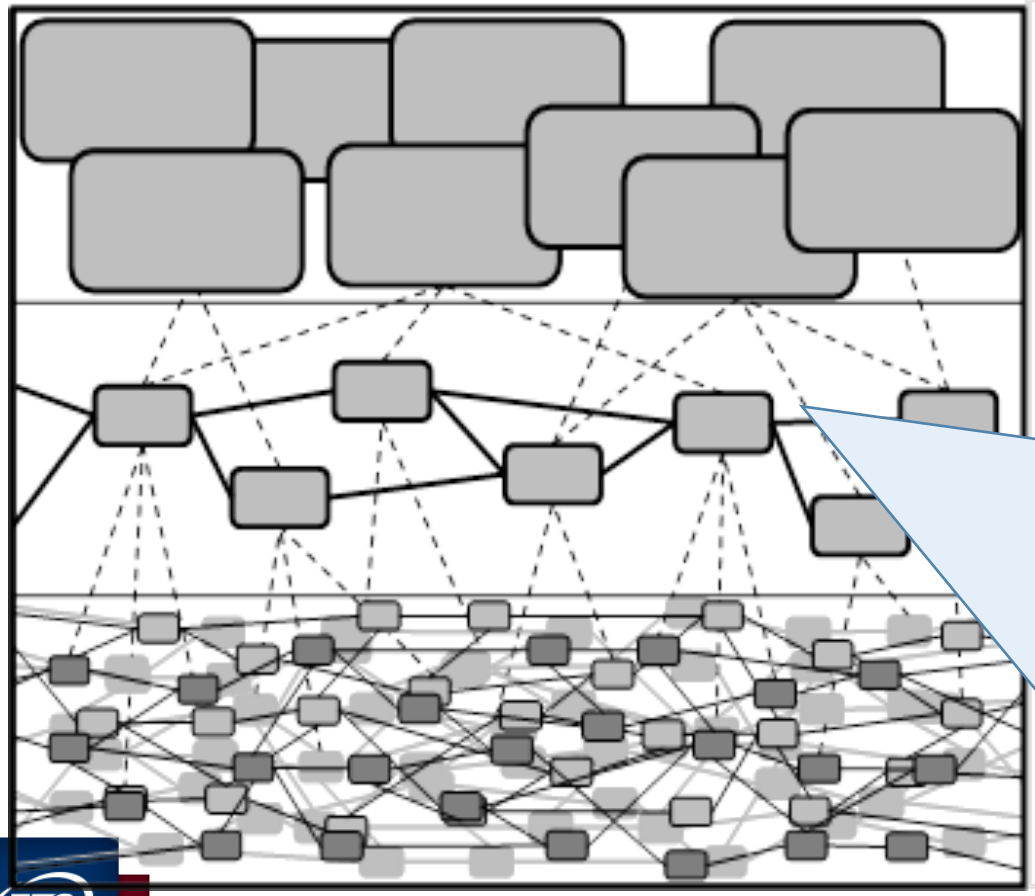
Snow & Lohman on "Measuring Traits"

The implication is that sign-trait interpretations of test scores and their intercorrelations are superficial summaries at best. At worst, they have misled scientists, and the public, into thinking of fundamental, fixed entities, measured in amounts.

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

Snow & Lohman, 1989, p. 317

A Situative/Sociocognitive Perspective

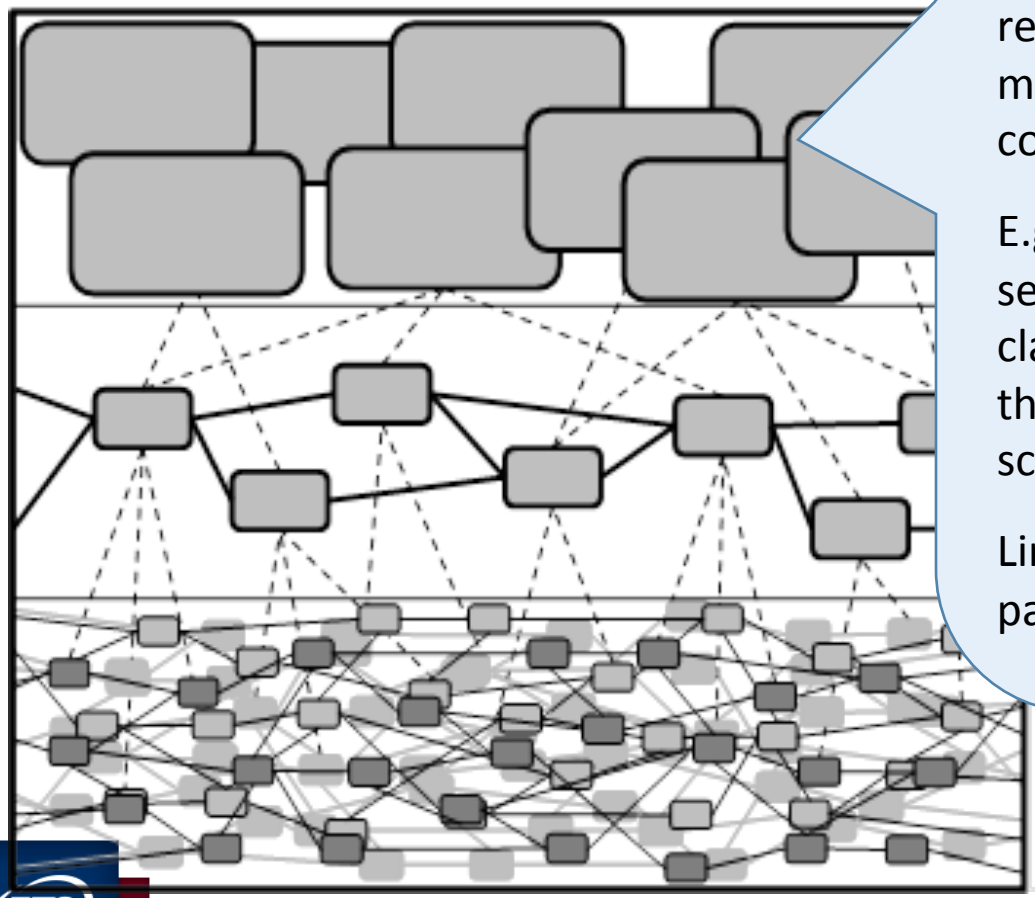


Human-level activity, persons acting within situations--the actions, events, and activities we experience as individuals.

We interact with the world and with each other: thinking, planning, conversing, reading, working, playing, solving problems, using representations, and cooperating or competing with family, friends, co-workers, etc.

Trait & behaviorist psychology are cast at this level.

A Situative/Sociocognitive Perspective



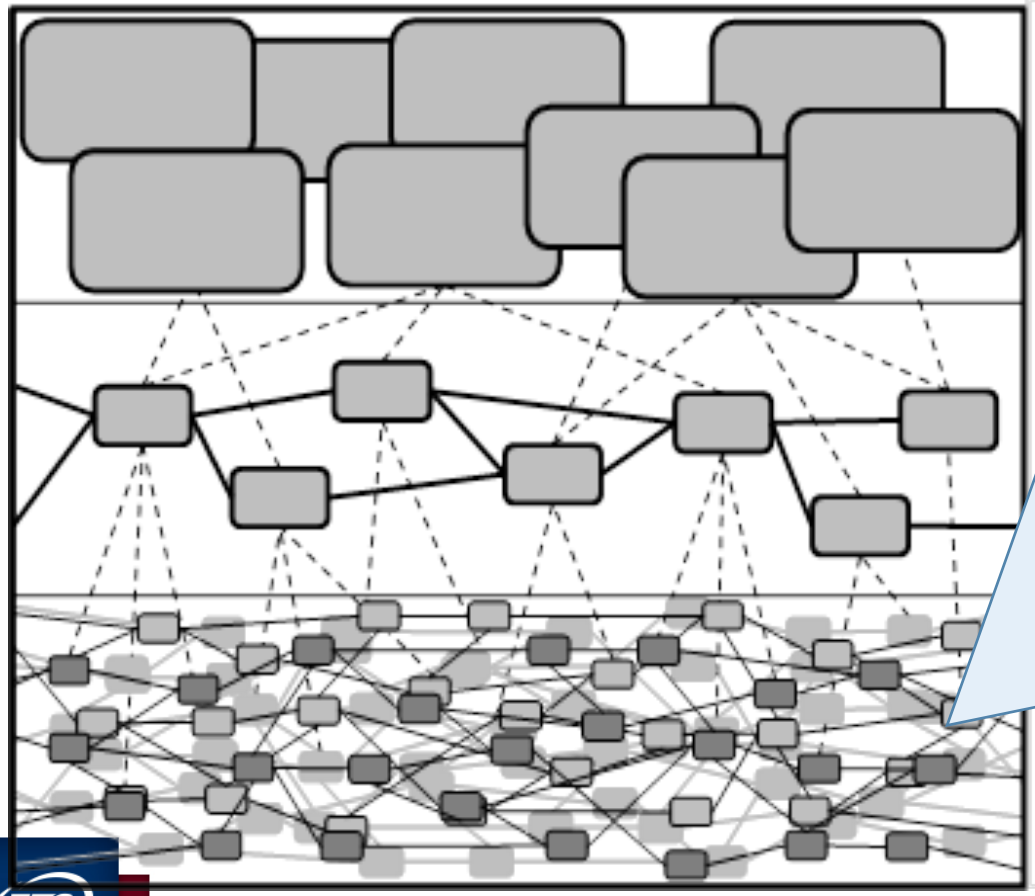
Between-persons patterns -- regularities in interactions of people in many overlapping identities, communities, affinity spaces

E.g., cultural models; language & semiotic systems; schemas for classrooms, offices, families; narrative themes; scientific models, arithmetic schemas.

Linguistic, cultural, substantive (LCS) patterns.

Milli-seconds	Within-individual events and processes that produce cognition	comprehension, learning, feeling

A Situative/Sociocognitive Perspective

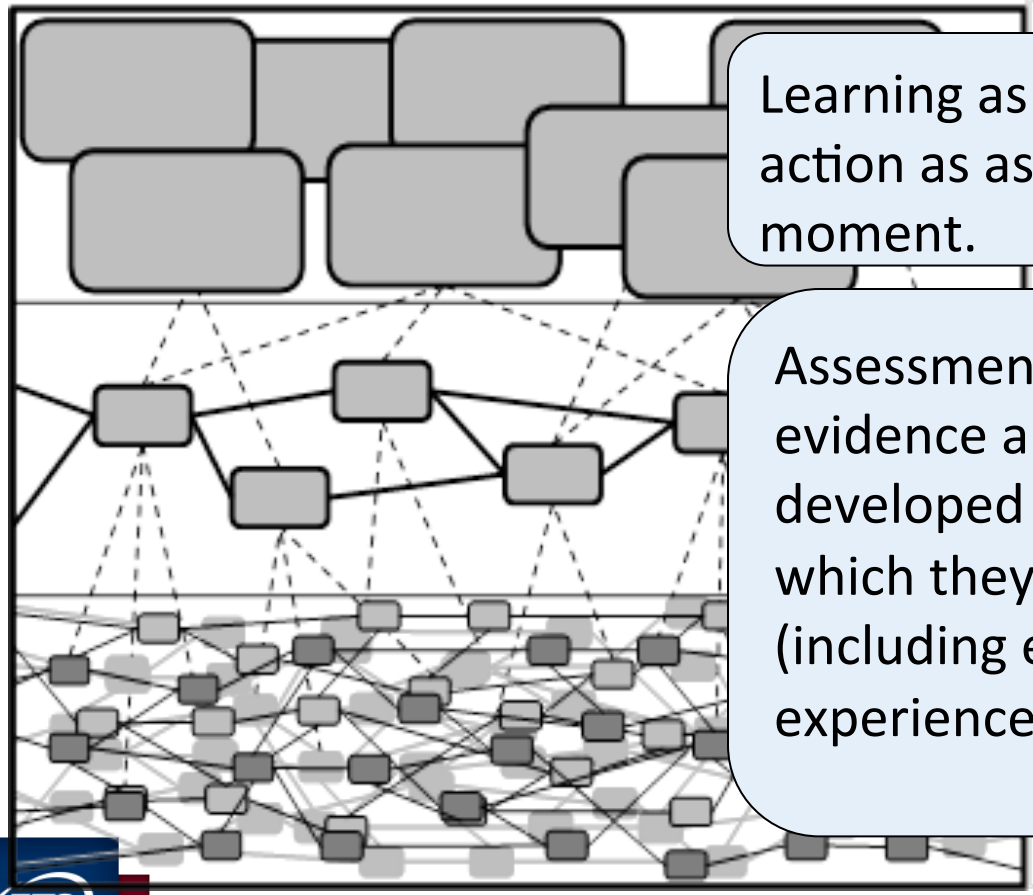


Within-person processes give rise to individuals' actions.

For successful human-level activity, neural activity patterns within individuals must both relate to LCS patterns and adapt to suit unique situations.

Young (2000) uses "*resources*" to refer to a person's capabilities to assemble particular patterns to understand, create, and act, in particular kinds of situations.

A Situative/Sociocognitive Perspective



Time Scale	Interaction	Examples
	Societal change:	Evolving
Weeks	through them	models
	produce cognition	

Learning as developing resources, action as assembling resources in the moment.

Assessment is evoking and interpreting evidence about resources people have developed and circumstances under which they can do what kinds of things (including extend them through further experiences)

Implications regarding **what** we assess (1)

Interest in persons' resources for more interactive, situated, activity.

- E.g: Being proficient in a language isn't just knowing a lot of words and grammar.
- "Communicative competence" as a construct?
- Pragmatic, strategic, metacognitive resources.
- Expectations, genres, purposes.

Implications regarding **what** we assess (2)

Capability is not just about individual cognitive resources …

- … for learning, for acting in the world, and for performing in assessments.
- Relevance of affective and conative factors.
- Key condition for learning.
- Key in interpreting assessment results.
- Key in usefulness of scores in prediction.
- Connects with cultural capital (LCS patterns, expectations, resources in a person's milieu)

Implications regarding **what** we assess (3)

Resources are developed in specific contexts, and are **not** well assessed in the same way and with the same interpretation for everybody.

- E.g., creativity, communication skills, troubleshooting, model-based reasoning.
- But these capabilities in particular domains – like troubleshooting hydraulics subsystems in the F-15 -- **are** much better approximated like this.
- And multiple contextualized experiences like this is how we develop more general resources.

Challenges and Opportunities re

Technology

Implications regarding **how** we assess (1)

With digital, interactive, environments, we can…

- Have interactive, continuous tasks (like SimCityEDU).
- Can include interactions with avatars or other humans.
- Simulate activity environments with high fidelity (like Hydrive)
- These can engage pragmatic aspects of capability.
- These can provide evidence about resources for interaction in environments.
- Also suited to learning, with engagement and situated contexts. (SimCityEDU is a formative assessment)

Implications regarding **how** we assess (2)

We can capture rich data…

- log files, video, physical, constructed products (like architectural designs).
- Interactions among multiple people (or many people, or massively many people – game telemetry).
- Use richer context to increase engagement.
- Monitor **states** of engagement and affect (not θs).

But how do we make sense of the data?

Challenges and Opportunities re

Psychometrics

1. Models

Data features beyond familiar test models in that the rich data:

- Continuous activity
- Conditional dependence
- Examinee actions change the situation
- Multiple proficiencies (θs).
- Different proficiency / observable combinations ($\mathbf{X}|\theta$)
- Changing proficiencies
- Multiple modalities
- Interaction among examinees

Basic ideas of psychometric models (1)

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.

Charlie Lewis, 1986.

Basic ideas of psychometric models (2)

The conditional probability model-fragments:

$$p\left(X_{ijk}\middle|\theta_i,\beta_j,\zeta_k\right)$$

where

X_{ijk} is the "observable" variable from the action(s) of "Person *i*" in "Situation *j*" given other relevant contextual variables *k*;

θ_i is the "proficiency" variable for "Person *i*" (might also subscript for time *t*);

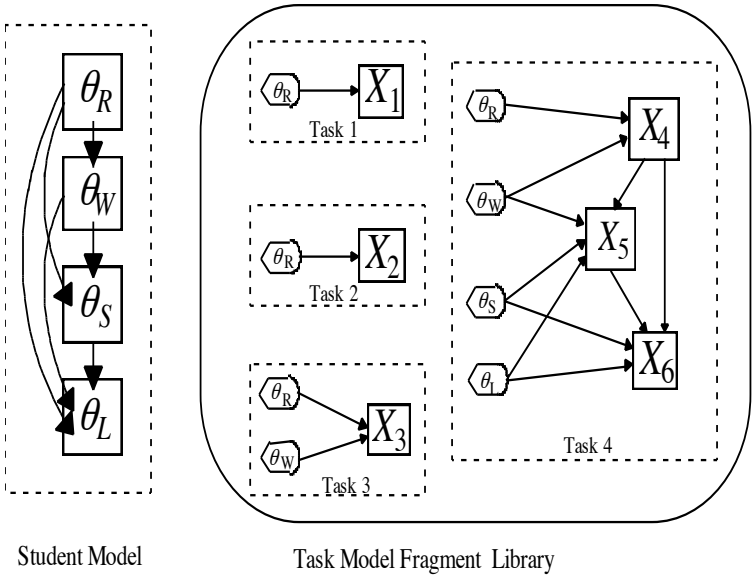
β_j is the effect of "Situation *j*"; and

ζ_k is the effect of other relevant contextual variables *k*.

Basic ideas of psychometric models (3)

Conditional independence:

$$p(\mathbf{X}|\theta,\beta,\zeta) = \prod_i \prod_j \prod_k p\left(X_{ijk}\middle|\theta_i,\beta_j,\zeta_k\right).$$



Can extend / modify in many directions as needed –

Data features beyond familiar test models in that the rich data:

- Continuous activity — Will say more later
- Conditional dependence
- Examinee actions change the situation
- Multiple proficiencies (θs).
- Different proficiency / observable combinations ($\mathbf{X}|\theta$)
- Changing proficiencies
- Multiple modalities
- Interaction among examinees

Data features beyond familiar test models in that the rich data:

- Continuous activity
- Conditional dependence
- Examinee actions change the situation
- Multiple proficiencies (θs).
- Different proficiency / observable combinations ($\mathbf{X}|\theta$)
- Changing proficiencies
- Multiple modalities
- Interaction among examinees

Model dependencies;
Markov processes (LaMar);
Condition on contextual factors

Data features beyond familiar test models in that the rich data:

- Continuous activity
- Conditional dependence
- Examinee actions change t
- Multiple proficiencies (θs).
- Different proficiency / observable combinations ($\mathbf{X}|\theta$)
- Changing proficiencies
- Multiple modalities
- Interaction among examinees

Lean hard on psychological theory and task/interface design

Data features beyond familiar test models in that the rich data:

- Continuous activity
- Conditional dependence
- Examinee actions change the situation
- Multiple proficiencies (θs).
- Different proficiency / observable combinations ($\mathbf{X}|\theta$)
- Changing proficiencies
- Multiple modalities
- Interaction among examinees

Dynamic Bayes nets; Bayesian knowledge tracing

Data features beyond familiar test models in that the rich data:

- Continuous activity
- Conditional dependence
- Examinee actions change the situation
- Multiple proficiencies (θs).
- Different proficiency / observable combinations ($\mathbf{X}|\theta$)
- Changing proficiencies
- Multiple modalities
- Interaction among examinees

Will say more later

Data features beyond familiar test models in that the rich data:

- Continuous activity
- Conditional dependence
- Examinee actions change the situation
- Multiple proficiencies (θs).
- Different proficiency / observable combinations ($\mathbf{X}|\theta$)
- Changing proficiencies
- Multiple modalities
- Interaction among examinees

Model dyads; condition on salient features of other actors; use avatars

Challenges and Opportunities re

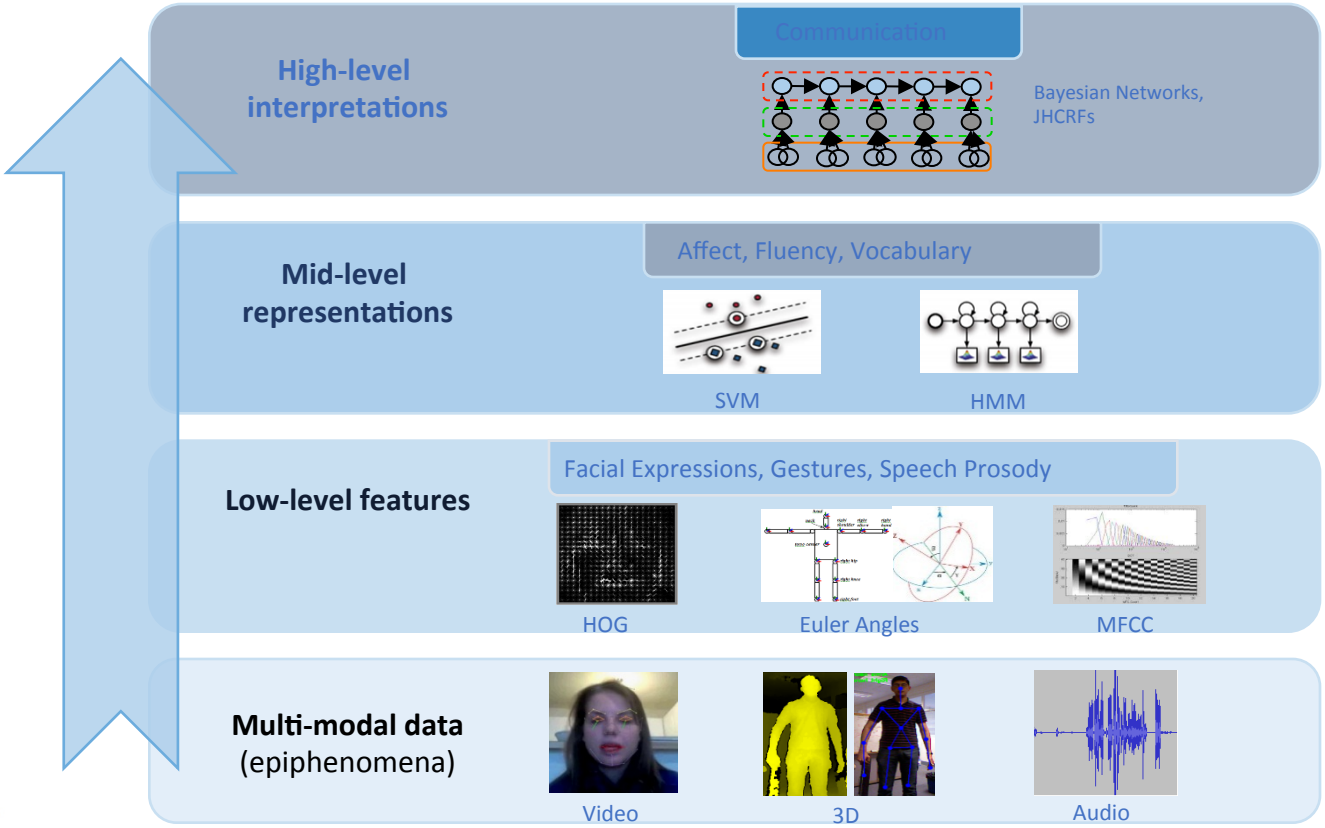
Psychometrics

2. Making Sense of Complex Performances

Hierarchical Inference

Khan & Kerr (2014, 2015)

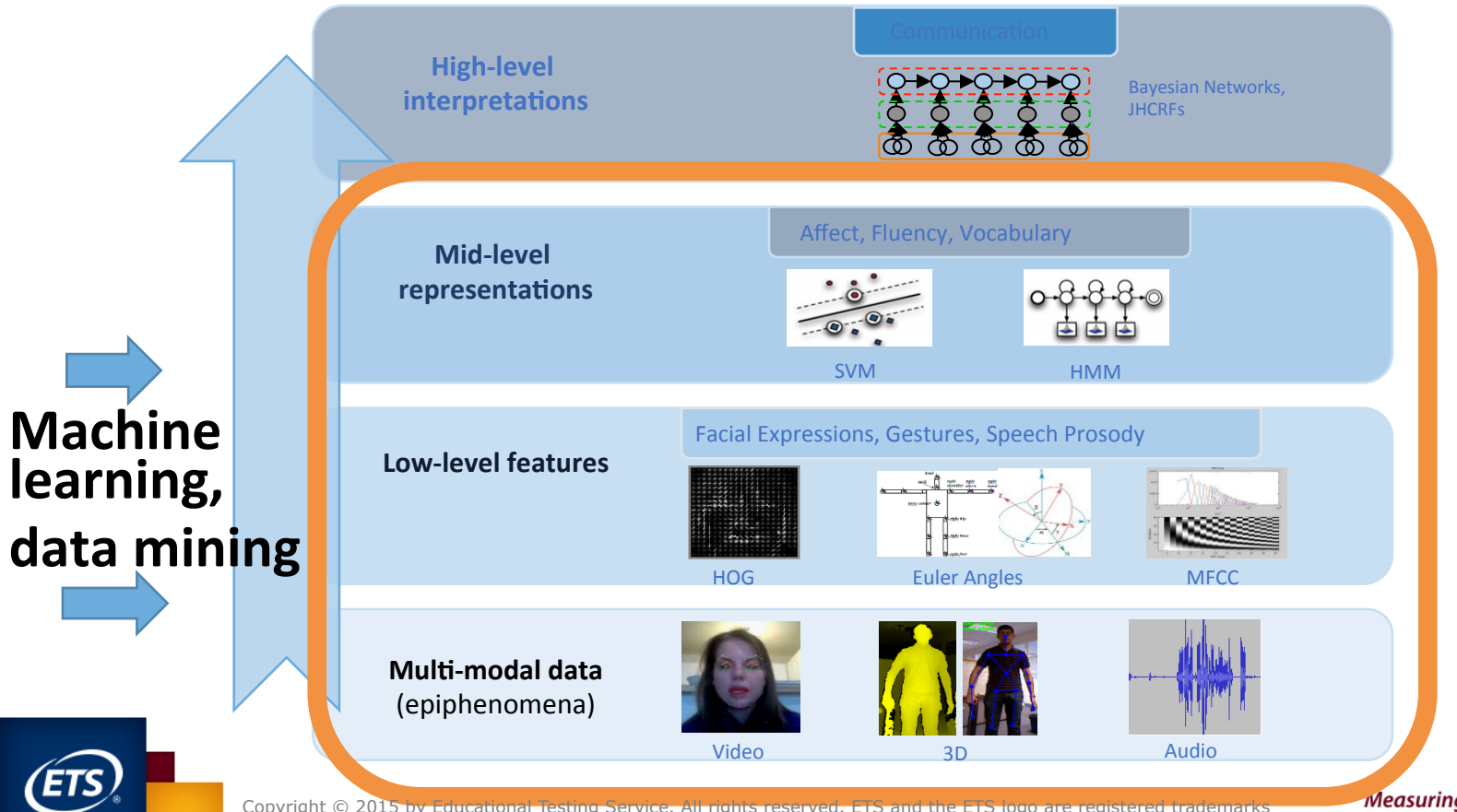
- Identify Constructs associated with behavioral patterns of interest (*Evidence*)
- Find evidence for these constructs from low-level multiple sensory data
  Hierarchies of evidentiary argument – can include up & down.



Hierarchical Inference

Khan & Kerr (2014, 2015)

- Identify Constructs associated with behavioral patterns of interest (*Evidence*)
- Find evidence for these constructs from low-level multiple sensory data
  Hierarchies of evidentiary argument – can include up & down.

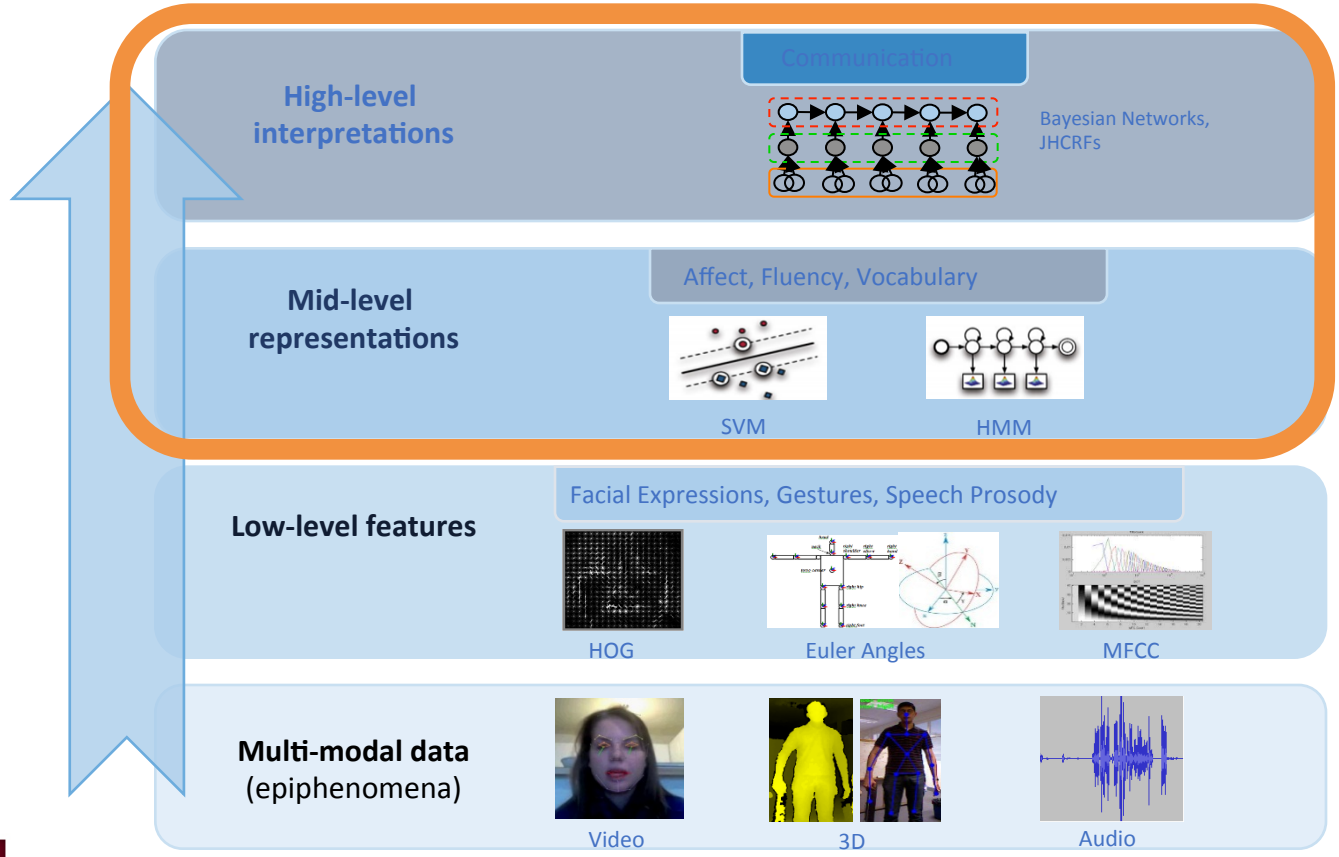


Hierarchical Inference

Khan & Kerr (2014, 2015)

- Identify Constructs associated with behavioral patterns of interest (*Evidence*)
- Find evidence for these constructs from low-level multiple sensory data

  Hierarchies of evidentiary argument – can include up & down.



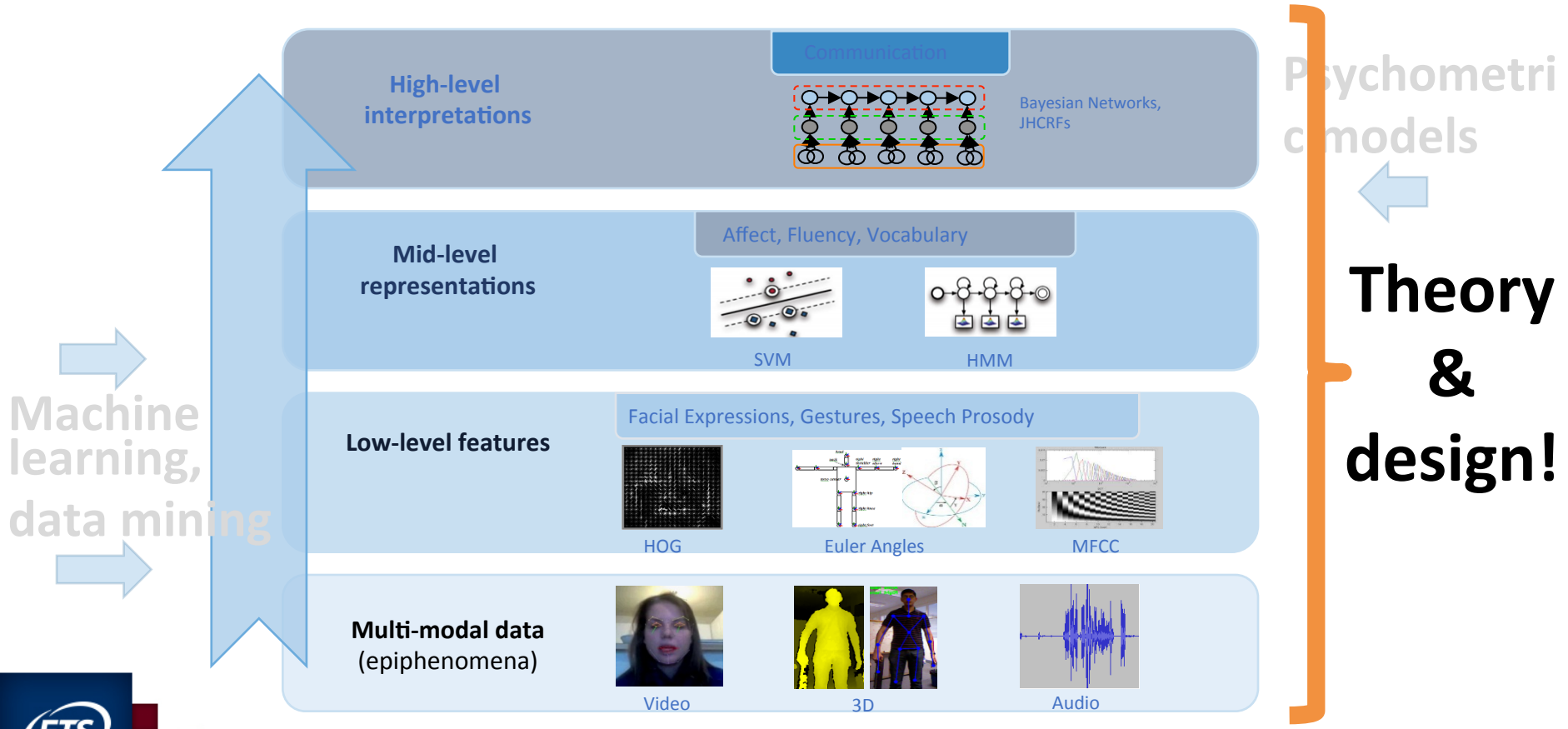
Psychometric models

Hierarchical Inference

Khan & Kerr (2014, 2015)

- Identify Constructs associated with behavioral patterns of interest (*Evidence*)
- Find evidence for these constructs from low-level multiple sensory data

  Hierarchies of evidentiary argument – can include up & down.

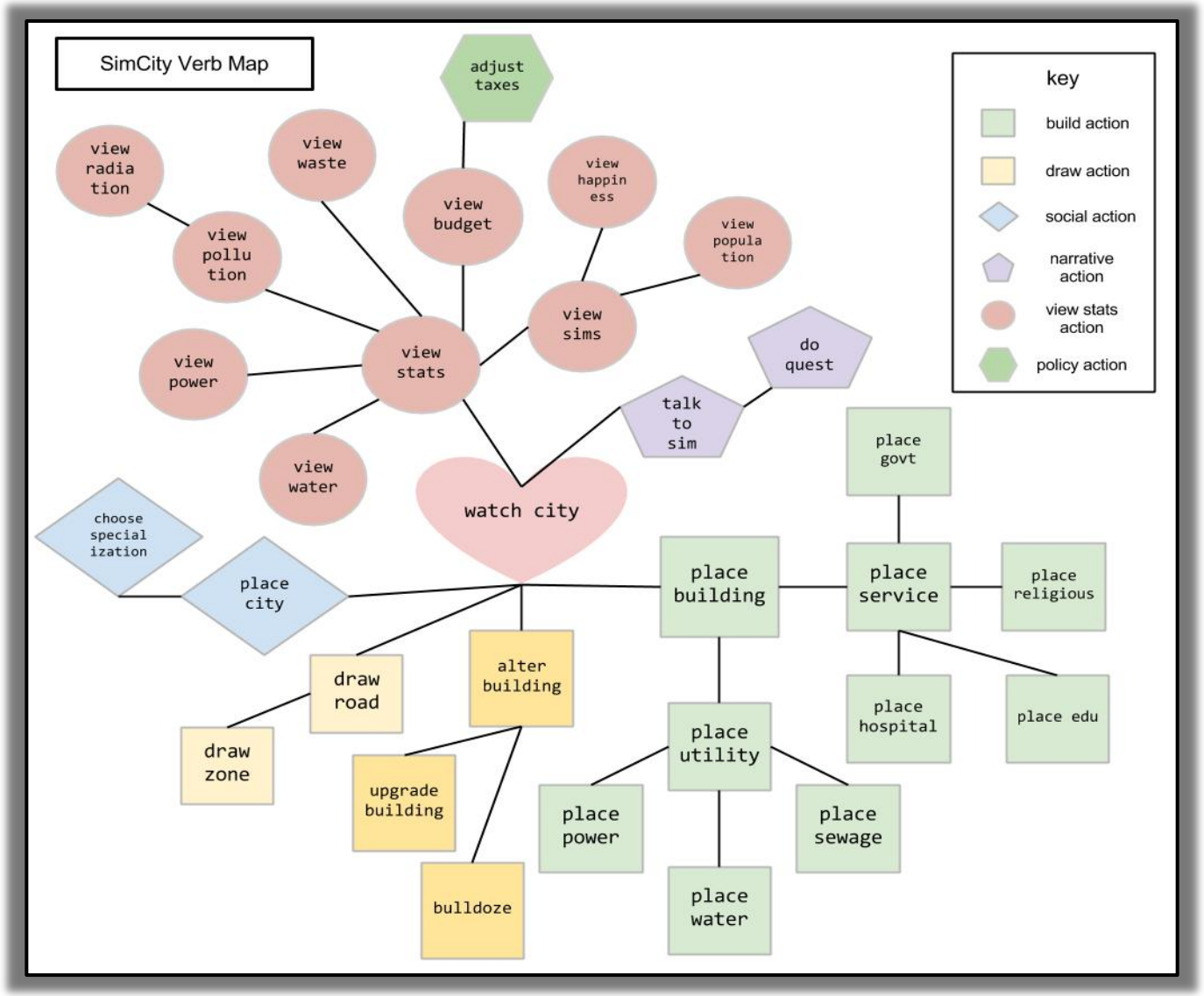


Quick Example

Event level log file: like in Pollution City

```
<stateInfo>
    <Init Fragment="600" Style="blue" Language="ENG" />
    <FSMStates Count="7">
        <finishButtonFSM>Stop Flashing</finishButtonFSM>
        <popupSubmitButtonFSM>Stop Flashing</popupSubmitButtonFSM>
        <submitButtonFSM>Stop Flashing</submitButtonFSM>
        <continueButtonFSM>Stop Flashing</continueButtonFSM>
        <nextButtonFSM>Stop Flashing</nextButtonFSM>
        <timeoutFSM>Clear</timeoutFSM>
        <playmakerFSM>Finish</playmakerFSM>
    </FSMStates>
    <OtherVars Count="3">
        <F6_Response>A</F6_Response>
        <F6_Reason1>asdf</F6_Reason1>
        <F6_Reason2>asdf</F6_Reason2>
    </OtherVars>
</stateInfo>
<itemResult accessionNumber="TestAccNum" itemType="SBT" childItemAccessionNumber="176"
blockCode="TestBlockCode">
    <responseVariable cardinality="single" baseType="string">
        <candidateResponse>
            <value><![CDATA[{"Selection of relevant questions":"Y,Y,Y,N"}]]></value>
        </candidateResponse>
    </responseVariable>
    <responseVariable cardinality="single" baseType="string">
        <candidateResponse>
            <value><![CDATA[{"How far away is the well?":"(a)Yes","Follow-up":"(a)There is probably not
enough water underground"}]]></value>
        </candidateResponse>
    </responseVariable>
    <responseVariable cardinality="single" baseType="string">
        <candidateResponse>
            <value><![CDATA[{"Wells in other villages?":"(b)No","Follow-up":null}]]></value>
        </candidateResponse>
    </responseVariable>
    <responseVariable cardinality="single" baseType="string">
        <candidateResponse>
```

Semantic Level: Pollution City Verb Map



Tactical Level: Pollution City

1. Identify instance of bulldozing old inefficient powerplant.
2. Identify sense-making antecedent actions:

   Was it after rezoning and plopping new efficient powerplant?

   or

   Was it before rezoning and plopping new efficient powerplant?

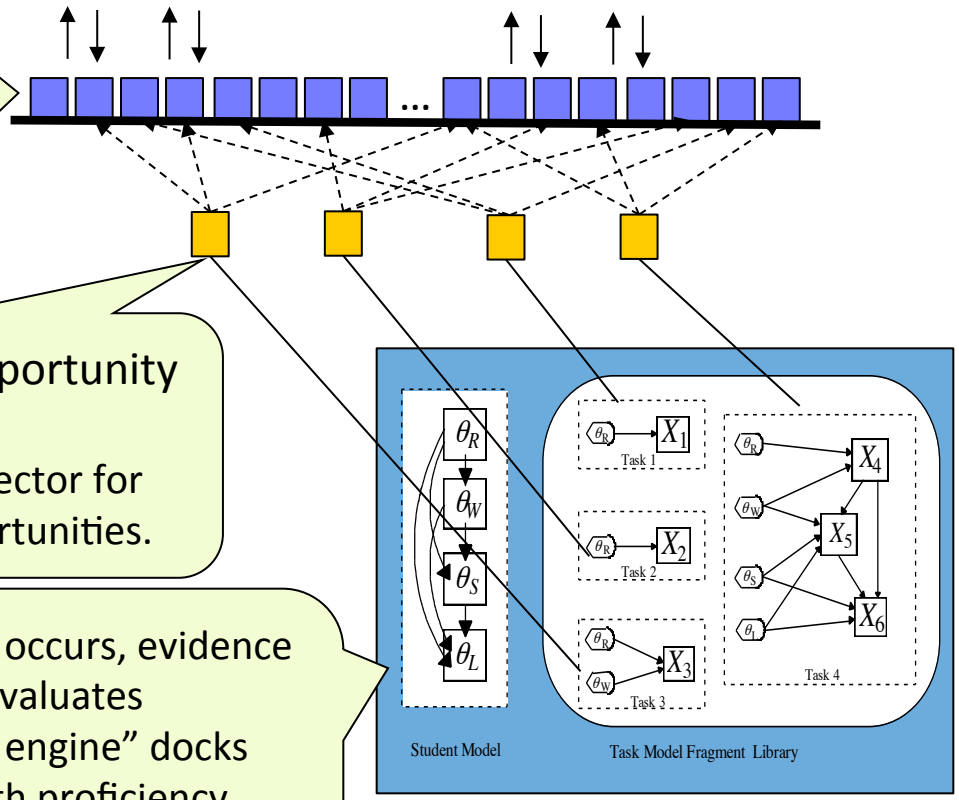
Psychometric Level: Pollution City

Flow of Activity

State vector.
Tracks relevant features of situations and past actions.

Evidence-bearing opportunity detectors.
Agents monitor state vector for evidence-bearing opportunities.

When a particular EBO occurs, evidence identification routine evaluates evidence, and "scoring engine" docks Bayes net fragment with proficiency model to update θ probability distribution.



Implications for the Five Questions

1. How predictive are elicited measurements of θ of both short-run and long-run outcomes?

Cisco's Packet Tracer and Hydrive simulation assessments are strongly contextualized. Studies have shown validity for learning and prediction.

These is a good deal of evidence for predictive validity of other contextualized simulator tasks – e.g., driving, patient management.

Less is known about decontextualized game and simulation assessments of higher level and non-cog proficiencies.

2. How important are incentives and contexts on measurements of θ?

Contexts are very important.

They make for better measurement for contextualized inferences, and worse measurement for pan-context inferences.

This is the finding of "low generalizability" in studies of performance assessment for broad skills like science investigation.

3. Do differences in environments change the predictive accuracy of elicited measures of θ?

I don't have much to say about this.

4. How can separate components of θ be identified?

My experience is not with trying to *identify* and *assess* components,

but rather using the interactive contexts to create situations where we can *condition* on high values of them so as to increase learning and validity of assessment with respect to cognitive components.

5. Are measurements of θ comparable across elicitation strategies?

There is reason to be skeptical given, e.g., …

- Sociocognitive research on initial (and often continuing) bonding of resources to learning conditions,
- e.g., studies such as those by Lave (supermarket math) and Saxe (candy-selling).
- Low generalizability results from Shavelson et al.
- Frederiksen's gunners' mate p&p vs hands-on assessments.

Conclusion

My conclusions* (1)

Relatively low ceiling on how much fancy technology & psychometrics can make noncog & trait SEMP measurement better.

Some improvement is definitely possible because interaction is possible. For SEMP testing, the best promise I see is in things like communication.

Ceiling due to contexuality of capabilities.

* My views do not necessarily represent official positions of ETS.

My conclusions* (2)

Relatively high potential for leveraging technology in contextualized assmt, as in SimCityEDU and Hydrive.

Relatively high potential for automated methods of identifying, synthesizing evidence and using psychometric models to at high levels of inferential hierarchies.

Best potential for learning and close prediction.

* My views do not necessarily represent official positions of ETS.

Thank you.

Some References

He, A. W. & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). Amsterdam/ Philadelphia: Benjamins.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R.J., Almond, R.G, & Lukas, J. (2004). A brief introduction to evidence-centered design. *CSE Technical Report.* Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. http://www.cse.ucla.edu/products/reports/r632.pdf

Mislevy, R.J., & Riconscente, M.M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Erlbaum. Online version available as *PADI Technical Report 9*. Menlo Park, CA: SRI International: http://padi.sri.com/downloads/TR9_ECD.pdf

Mislevy, R.J., Corrigan, S., Oranje, A., DiCerbo, K., John, M., Bauer, M.I., Hoffman, E., von Davier, A.A., Hao, J. (2014). *Psychometric considerations in game-based assessment*. New York: Institute of Play. http://www.instituteofplay.org/work/projects/glasslab-research/

Mislevy, R.J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine* (special issue on simulation), *178*(10) Supp., 107-114. http://www.cse.ucla.edu/products/reports/R800.pdf

Mislevy, R.J., Behrens, J.T., DiCerbo, K., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining, 4,* 11-48. http://www.educationaldatamining.org/JEDM/images/articles/vol4/issue1/MislevyEtAlVol4Issue1P11_48.pdf

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching, 33*, 1045-1063.

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