Evidence-centered Classroom Assessment Design

Russell Almond

Educational Psychology and Learning Systems
College of Education
Florida State University

Talk at AIR

Measuring the Wind

- The wind cannot be seen.
- But we can see *evidence* of wind.
- We can use this to build an instrument to measure wind: *an anemometer*.

Messick Quote

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? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics.

(Messick, 1994, p. 16)

Four Questions

- What are we measuring? — Competencies and Constructs
- How are we measuring? — Evidence
- Where are we measuring? — Task Contexts
- How much are we measuring? — Measurement/Instruction Plan

Evidential Reasoning

- Claims and Evidence / Toulmin Diagram
- Bayesian Induction – Schum (1964) *The Evidential Foundations of Probabilistic Reasoning*
- Weights of Evidence
- Engineering Assessments for Good Evidence

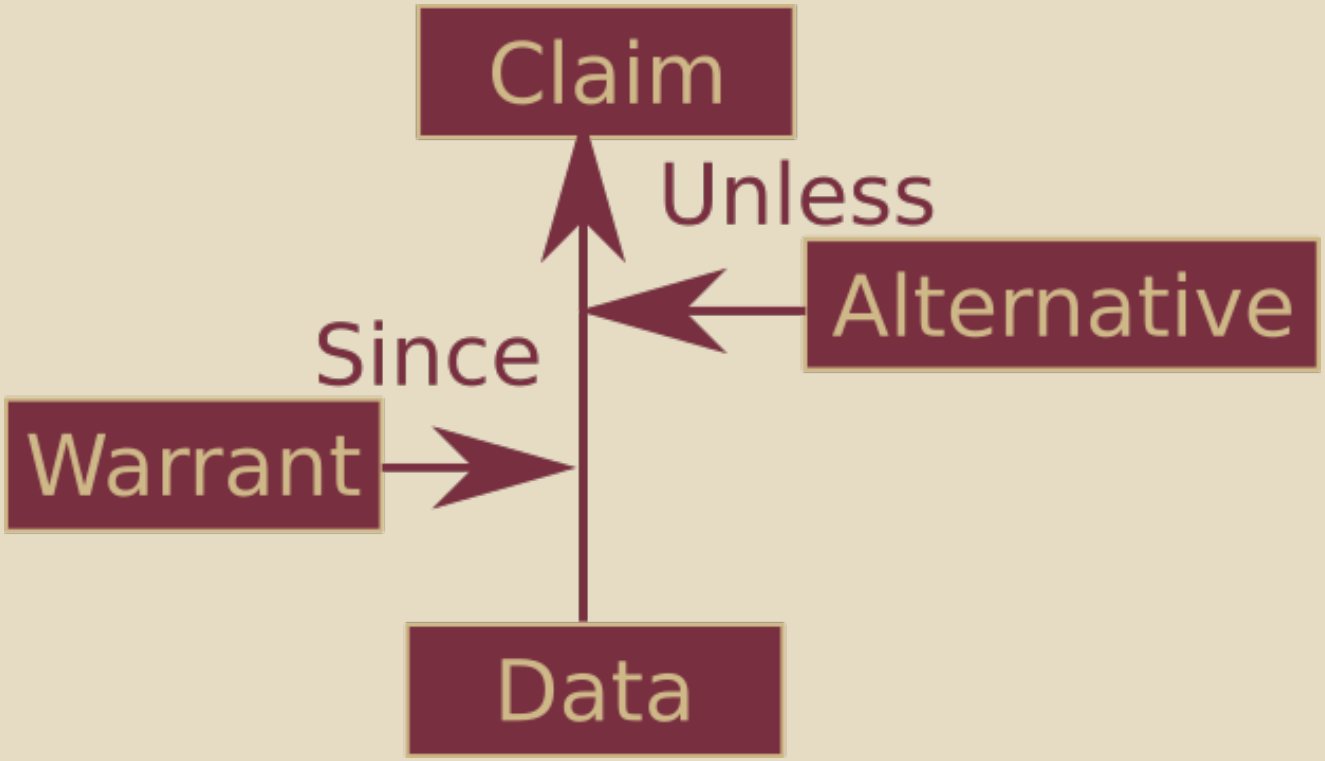
Claims

- A precise statement we wish to make about a learner
- Must pass the *clarity test* — could an observer with infinite evidence unambiguously decide if claim holds.
- Often hierarchical
- Related to standards

Evidence

- Category of observation which is more likely when claim holds
- *Data become Evidence when linked to a claim.*
- Grouping of possible work products from a task.
  - Correct or Incorrect
  - Contains or does not contain key idea
  - Applied or did not apply particular procedure
  - Well developed thesis or weakly developed thesis or no thesis or off topic.
- Evidentiary value may depend on context

Toulmin Diagram



- *Claim:* What we want to establish about learner.
- *Data:* What we observe about learner's performance.
- *Warrant:* Why we believe about learner's performance is related to claim.
- *Alternative:* explanation for learner's performance unrelated to claim.

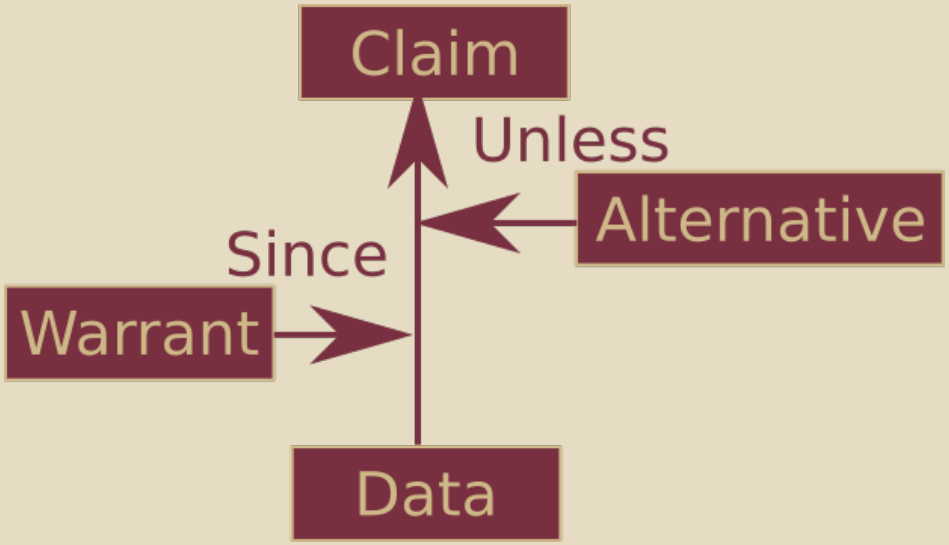
Warrants

- Reasons why we believe that evidence is more likely when claim holds.
- Reasons why we believe that evidence is less likely when claim does not hold.
- Validity evidence
- Design task contexts to maximize warrants

Rebuttals

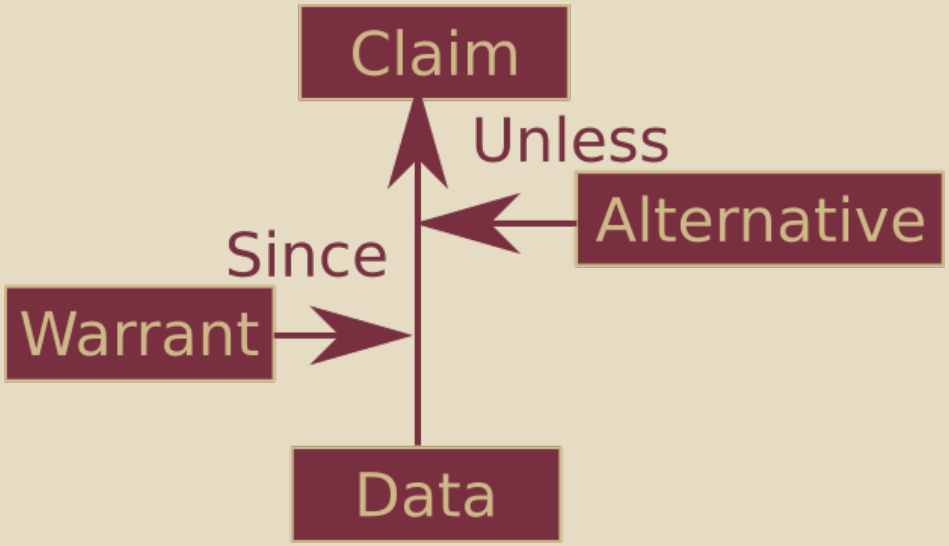
- Reasons why evidence might appear when claim does not hold.
- Reasons why evidence might not appear even when claim does hold.
- Alternative explanations for success/failure.
- Construct irrelevant skills!

Accessibility Example 1



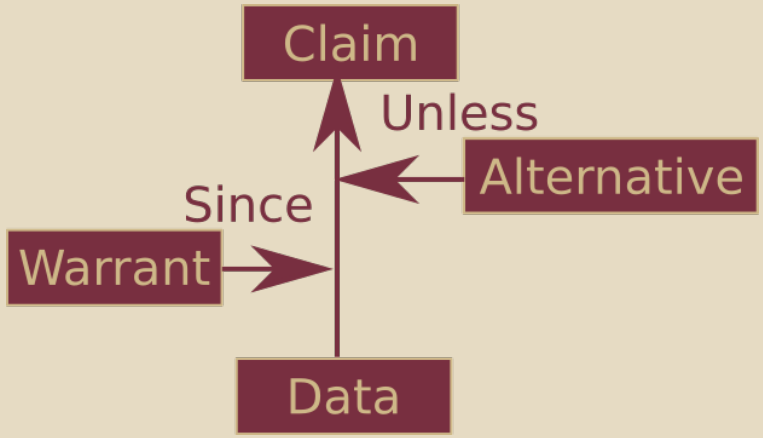
- *Claim:* Learner can decode English.
- *Data:* Learner successfully read pseudo-word list
- *Warrant:* Pseudo-words are defined with regular English pronunciation.
- *Alternative:* Learner has low vision and print is small.

Accessibility Example 2



- *Claim:* Learner can decode English words.
- *Data:* Learner successfully read pseudo-word list
- *Warrant:* Pseudo-words are defined with regular English pronunciation.
- *Alternative:* Read aloud accommodation was available.

Exercise: Make a Toulmin Diagram



- *Claim:*
- *Data:*
- *Warrant:*
- *Alternative:*

Log-odds as a Measure of Confidence

- Let C represent the claim holding: $\overline{C}$ claim not holding.
- Let $\Pr(C)$ represent the probability of the claim holding (for a given individual)
- Consider $\log \frac{\Pr(C)}{\Pr(\overline{C})}$:
  - Positive—claim more likely true than false
  - Zero—claim true and false equally likely
  - Negative—claim more likely false than true

The Effect of Evidence

- Let E represent a (collection) of evidence
- $\Pr(C|E)$ and $\Pr(\overline{C}|E)$ are updated probabilities of claim (*given* evidence)
- New log-odds $\log \frac{\Pr(C|E)}{\Pr(\overline{C}|E)}$
- Measure of confidence is *subjective* in the sense that two raters with access to different collections of evidence will come to different conclusions.
- Measure of confidence is *objective* when raters with the access to the same collection of evidence will come to the same conclusions.

The Weight of Evidence

- Consider how much log-odds changes after observing evidence

$$\log \frac{\Pr(C|E)}{\Pr(\overline{C}|E)} - \log \frac{\Pr(C)}{\Pr(\overline{C})} \tag{1}$$

- By Bayes theorem this equals

$$W(C:E) = \log \frac{\Pr(E|C)}{\Pr(E|\overline{C})} \tag{2}$$

- This is called the *weight of evidence* (WOE)

Good Evidence

- Weight of Evidence

$$W(C:E) = \log \frac{\Pr(E|C)}{\Pr(E|\overline{C})} \tag{3}$$

- $\Pr(E|C)$ is probability of seeing evidence when claim holds
- $\Pr(E|\overline{C})$ is probability of seeing evidence when claim does not hold
- E is good evidence if $\Pr(E|C)$ is high and $\Pr(E|\overline{C})$ is low

Tasks that are too hard

- Consider a task for which both $\Pr(E|C)$ and $\Pr(E|\overline{C})$ are low.
- The ratio will be small, so evidence will be small.
- Conclusion: tasks that are two hard produce poor evidence (especially if learner is unsuccessful).
- Need to differentiate between two types of difficulty:
  - *Game difficulty*—difficult because of construct irrelevant skills that make it harder.
  - *Psychometric difficulty*—difficult because it requires more of measured construct to produce evidence

Tasks that are too easy

- Consider a task for which both $\Pr(E|C)$ and $\Pr(E|\overline{C})$ are low.
- The ratio will be small, so evidence will be small
- Conclusion: tasks that are two easy produce poor evidence (especially if learner is unsuccessful)
- However, easy tasks are often useful pedagogically:
  - Build learner self-efficacy and self-esteem
  - Provide asymmetric evidence: better at identifying learners who have fallen behind growth targets

Evidence Chaining

- Conditional Weight of Evidence: Already seen E_1, now see E_2

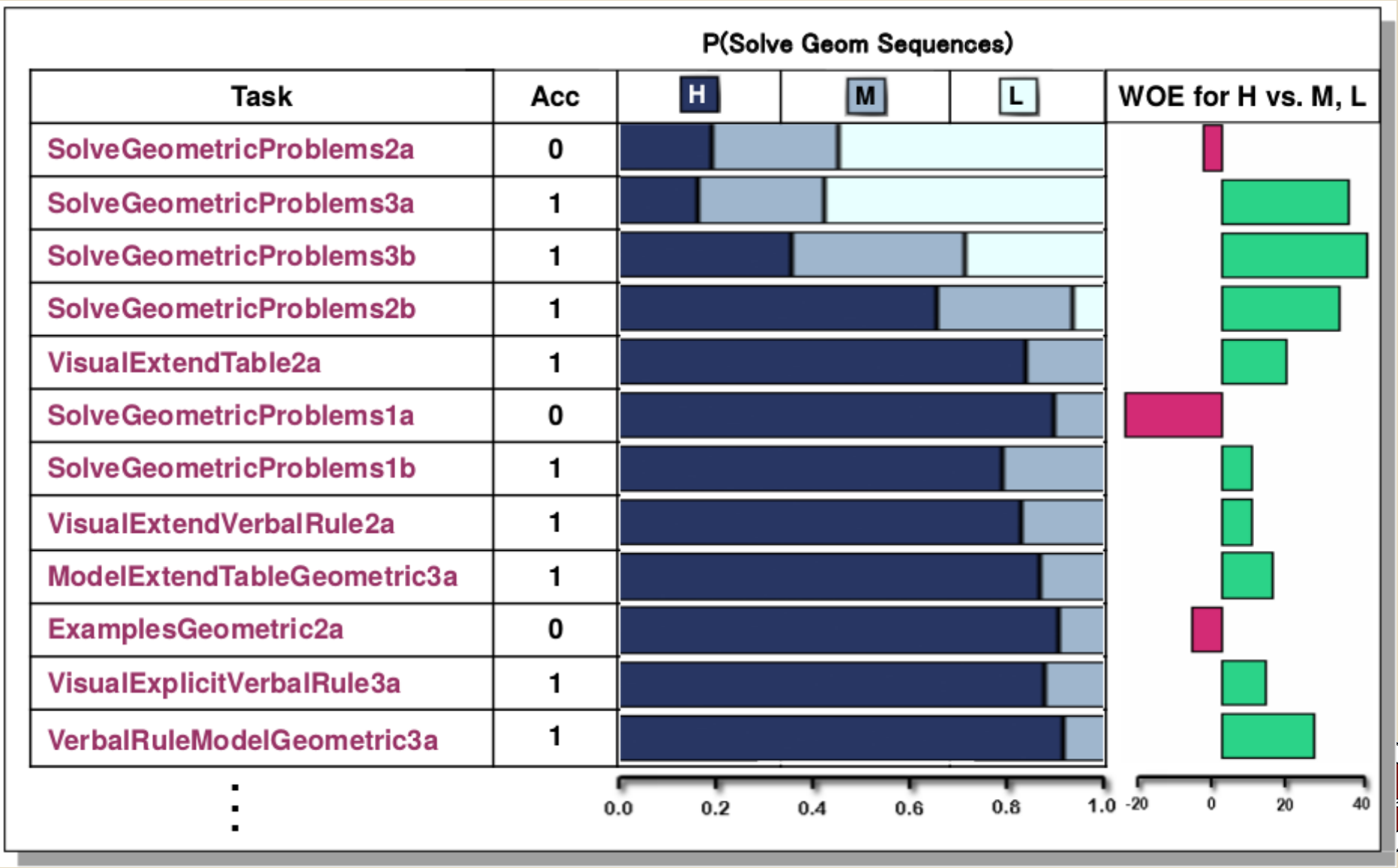
$$W(H : E_2|E_1) = \log \frac{\Pr(E_2|C, E_1)}{\Pr(E_2|\overline{C}, E_1)} \tag{4}$$

- Additive

$$W(H : E_2, E_1) = W(H : E_2|E_1) + W(H : E_1) \tag{5}$$
$$= W(H : E_1|E_2) + W(H : E_2) \tag{6}$$

- Order sensitive (Decreasing evidence from tasks of the same type).

Evidence Balance Sheet



Expected Weight of Evidence

- Expected weight of evidence (EWOE) is average (expectation) over possible evidence from task/activity

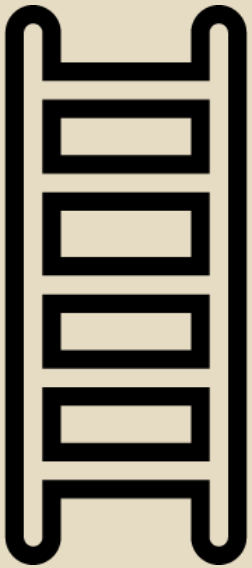
$$EW(C:E) = \sum_{j=1}^{n} W(C|e_j) \Pr(e_j|C) \tag{7}$$

- Task that maximizes EWOE is good choice for next.
- Appears to be a good choice from learning perspective, too. (Shute, Hansen & Almond, 2008).
- Greedy search is not always optimal: Combination of two or more tasks might be better.

Four Elements of Assessment Design

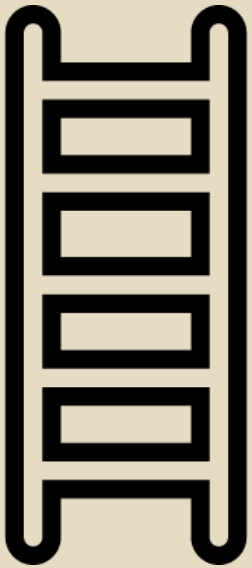
- Competencies and Constructs
- Evidence
- Assessment Contexts
- Assessment/Learning Plan

Larry's Ladder (Larry Ludlow)



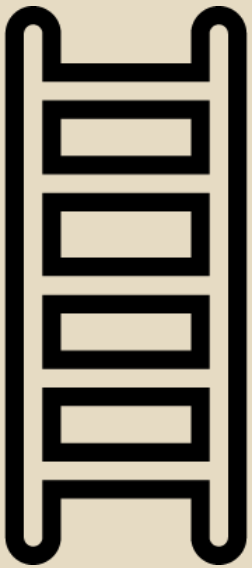
- A group of related claims are stacked to make a *competency*
  - Carry out operations with mixed numbers
  - Paragraph level writing
- Also could be non-academic construct
  - Self-regulation
  - Satisfaction
- Generally higher is better
- Goal is to figure out how far up the ladder learners are (and how to get them higher)

Russell's Rungs



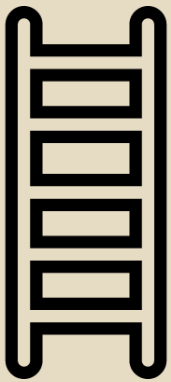
- *Each rung on the ladder should represent a difference of at least one claim.*
- Claims (and standard) are placed in various places in the ladder
- Learning progressions are a natural place to start.
- Establish a scale by looking at High, Medium and Low points

Three Views of Competency



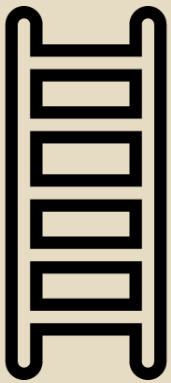
- Descriptions of People
- Descriptions of Evidence
- Description of Task Contexts
- Establish a scale by looking at High, Medium and Low points

Ladder Example



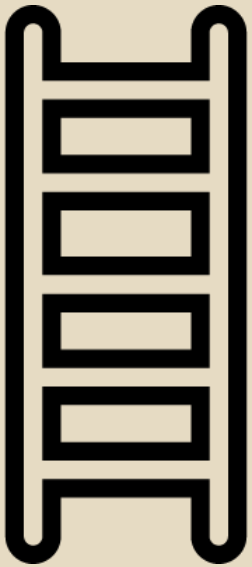
Height	Person	Evidence	Task
Liberal	Favors restricting guns	*Agrees* with proposal for stricter control	Agrees with proposal for *stricter control*
Moderate	Favors no change	*Indifferent* to proposal for stricter control	Agrees with proposal for *no change*
Conservative	Favors making guns more available	*Disagrees* with proposal for stricter control	Agrees with proposal for *more access*

Ladder Exercise



Height	Person	Evidence	Task
High			
Medium			
Low			

Evidence: Data linked to a Ladder



- *Observations become evidence when they are placed on the ladder.*
- Raw work products from tasks are categorized
- Each category is placed on the ladder.
- *Rules of evidence* (Rubrics) describe categories.
- *Weights of evidence* describe how far up or down ladder to move the estimate.

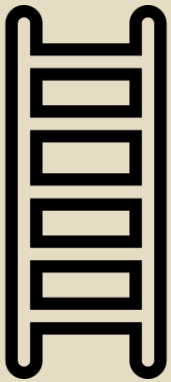
Rubrics: Rules of Evidence

- Start with work product
  - Selection on multiple choice test
  - Numeric answer on math problem
  - Essay
  - Oral presentation
  - Observations made during group work
- Sort work products into piles (categories)
- Write down features which separate work in different piles
- Associate categories with rungs on the ladder
  - If two categories map to the same rung, they can be merged

Weights of Evidence

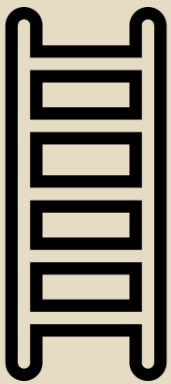
- Assign point values to categories, more points for being higher on the ladder
  - If there is more than one ladder, may be multiple weights
- Ideally, based on psychometric difficulty, not game difficulty.
  - Common practice of assigning weight on amount of effort is not ideal.

Evidence Worksheet



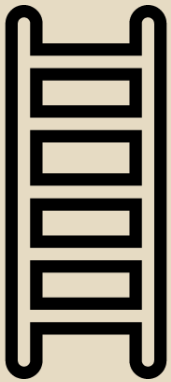
Category	Points	Features
Strong		
Moderate		
Weak		

Evidence Worksheet: Example



Category	Points	Features
Correct Answer and Explanation	Physics+2, Explanation+1	Prediction is correct, and explanation references Newton's laws of motion
Correct Answer, Incomplete Explanation	Physics +1, Explanation -1	Prediction is correct, but explanation is incorrect or incomplete
Incorrect Answer	Physics-1, Explanation =0	Prediction is missing or not correct

Evidence Worksheet: Exercise



Category	Points	Features
Strong		
Moderate		
Weak		

Assessment Contexts

- Classical items and item sets
- Extended Constructed response tasks
- Observations in the middle of larger activities
- Work product is determined by the item context
- Features of the context may determine how far up the ladder the task is

Context Features

- Features related to source material
  - Difficulty and length of source texts
  - Number of digits in math problems
  - Number of steps required for solution
  - Working memory load — irrelevant details
- Features related to available tools
  - Calculators
  - Dictionaries
  - Open Book/Internet
  - Individual or Team work
- Features related to possible answers
  - Expected answer form
  - Plausibility of distractors for selected-answer tasks
  - Scaffolding of answer style

Incidental Features

- Features that do not change difficulty or evidentiary focus are *incidentals*
  - Name of actors in story (within reason)
  - Exact numeric values in problems (within a restricted range)
- Put new clothing on task shells (automatically generated tasks)
- Watch out for incidentals which are not incidental

Features that drive difficulty

- Features that do change difficulty (or evidentiary focus) are *radicals*
  - Complexity or reading passage
  - Number of digits in math problem
- Drive task to provide evidence about different rungs on the ladder
- Watch out for tasks that drive *game difficulty* instead of *psychometric* difficulty
- This can be done with Likert-type agree/disagree items:
  - I *always* start studying *well before* the date of the exam.
  - I *usually* start studying *the night before* the exam.

Optimal Difficulty

- Tasks that are too hard or too easy for target population provide little evidence.
- Best evidence is usually around the point where learner has 50-50 chance of producing evidence
- In the middle of Vygotsky's zone of proximal development — Between Falmange's inner and outer fringes
- Tasks which are optimal for measurement are also optimal for learning
- *Every assessment can be formative if learners are given proper feedback.*

Features that change evidentiary focus

- Some task features may change task so much that it provides evidence for a different competency
  - Change number in math work problem to algebraic expressions
  - Change prompt on writing task from "summarize source" to "agree or disagree"
  - Add "Explain your reasoning"
- Sometimes want to manipulate these features to span a number of constructs
- Sometimes want to control these features to not get off topic

Features that add construct irrelevant variance

- Some features make task harder even for people with the target skill.
  - Large number of digits in arithmetic problem
  - Unfamiliar vocabulary not related to construct
  - Negative or complicated wording of instructions
  - Lots of background distraction
- These features lower discrimination (evidentiary value) of task
- Beware of unmodeled/unmeasured skills
- This can rise to a fairness issue if some members of target population have the skill and some do not

Limited testing resources

- Two types of tasks:
  - *Natural Tasks*—Valued work, unconstrained by time and resources
  - *Assessment Tasks*—Take into consideration constraints on time (learner and instructor), material, scoring speed, etc.
- Start with natural task and try to understand evidence and important features.
- To what extent can these be reconstructed in classroom
- Especially important when designing simulators: focus on the right evidence.

Task Worksheet

Task Worksheet: Example

Task Worksheet: Exercise

Plans

- Plan Parts
  - Pool of Assessment activities
  - Pool of Instructional activities
  - Selection and Sequencing Rules
  - Stopping Rules
- Plans for different time ranges
  - Single test or worksheet
  - Class period
  - Class week
  - Unit or Chapter
  - Semester

Implicit variable definitions

- Competency variables (ladders) are defined by what tasks are used to measure them.
- Mismatch between claims and tasks ⇒ Incorrect interpretation
- Consider a ladder labeled "Understands tables and graphs"
  - But a test with only graph tasks
- This is key for validity of assessment

Spanning contexts

- Need to make sure all contexts are covered.
- Might need to sample from contexts if there is too much to cover.
- Need to plan across multiple contexts.
- Need to make sure there are adequate activities in the pool to cover what is needed.
- A feature matrix—Rows are tasks, Columns are features—helps achieve balance

Sufficient Evidence

- How many tasks of each type are enough?
  - Answer depends on purpose
  - The higher the stakes the more evidence is needed.
- Item Response Theory (and other psychometric models) can give a rigorous answer
- Heuristic: 6–10 per construct when high reliability is needed.
- A Q-matrix—Rows are tasks, Columns are constructs—helps achieve balance

Plan Worksheet

Plan Worksheet:Example

Plan Worksheet:Exercise

Its only a model!

- Competencies and Constructs
- Evidence
- Assessment Contexts
- Assessment/Learning Plan
- All of these are probably wrong.
- Look at data to see if they can be improved.

Difficulty and evidence

- Look at difficulty of items (proportion correct)
- Tasks that are too hard (low proportion correct)
  - Low evidentiary value
  - Negative effect on learner self-efficacy
- Tasks that are too easy (high proportion correct)
  - Low evidentiary value
  - Positive effect on learner self-efficacy
  - Useful for wash back effect on learner behavior
- Beware of different difficulties for different groups of learners

Discrimination and evidence

- Look at discrimination (correlation between task-level score and overall score)
- Low correlation indicates low evidentiary value
- Look for confusing wording
- Look for unmeasured skills required to solve item
- Look for game difficulty instead of psychometric difficulty

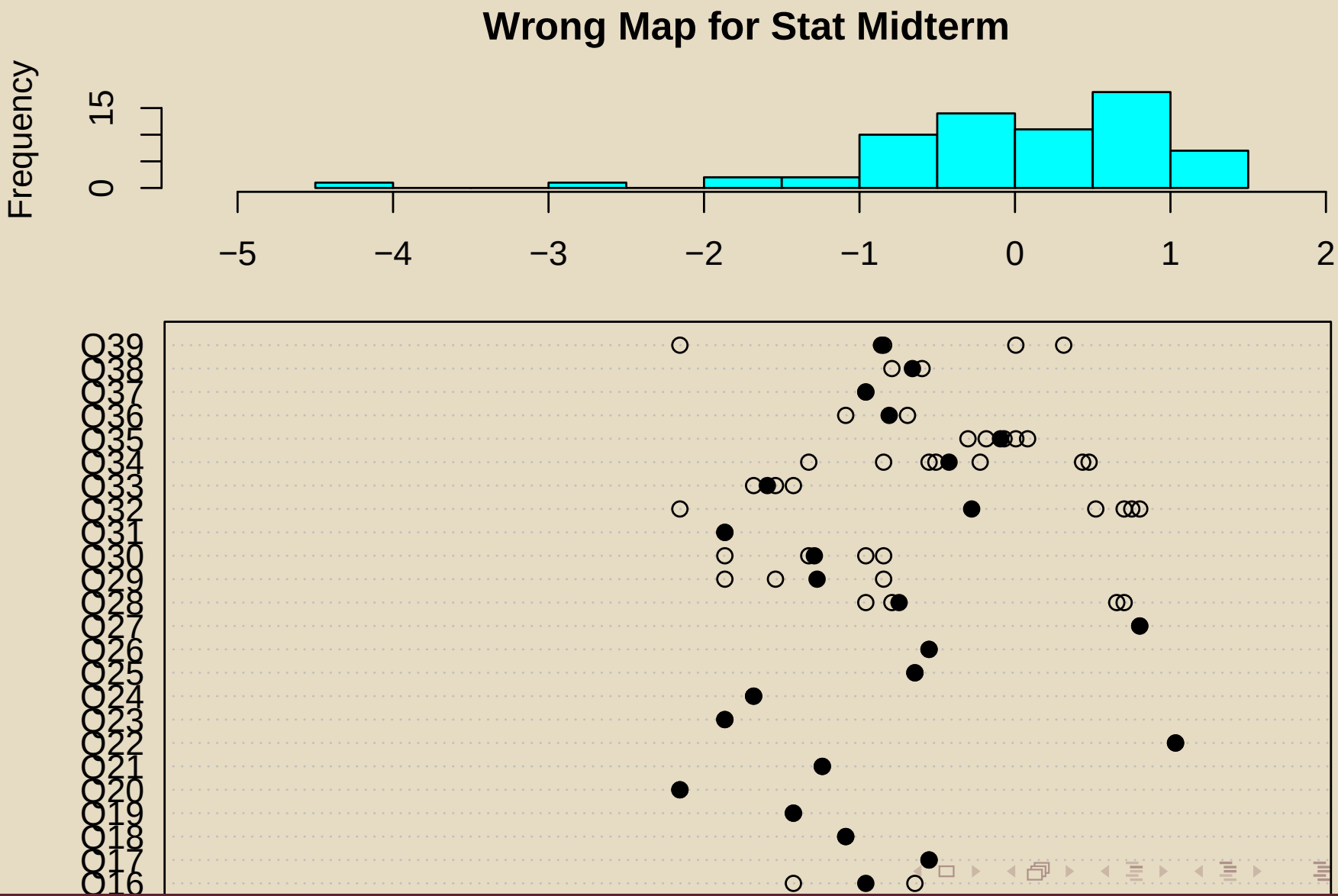
Reliability and plans

- Is there enough evidence to meet the purpose?
- Don't rely on a single source of evidence for each ladder — the more evidence the higher the reliability
- Are the tasks at the right places on the ladder?
- Do they cover all the ladders?
- Do the contexts span the construct? (Validity)

Two tools for checking ladders

- Wright Map
  - Scale tasks and people using the Rasch model
  - Plot them on the same scale
  - Do the items span the ability distribution?
- Wrong Map
  - If there are not enough data to do Rasch scaling
  - Scale people using a z-score transformation
  - Scale item difficulties using an inverse normal (probit) transformation.
  - Plot on same scale.

Wrong Map Example: Stat Class Midterm



Validity and Natural Tasks

- Go back to the natural tasks
  - Valued work
  - Too costly to measure (at scale)
- In the context of a smaller validity study these become fodder for validity studies
- With sufficient data link to *market basket* of ideal tasks.

Everybody's doing it.

- The following elements are in any assessment design
  - Competencies and Constructs
  - Evidence
  - Assessment Contexts
  - Assessment/Learning Plan
- They may have different names
- Explicitly naming the pieces supports reasoning about assessments