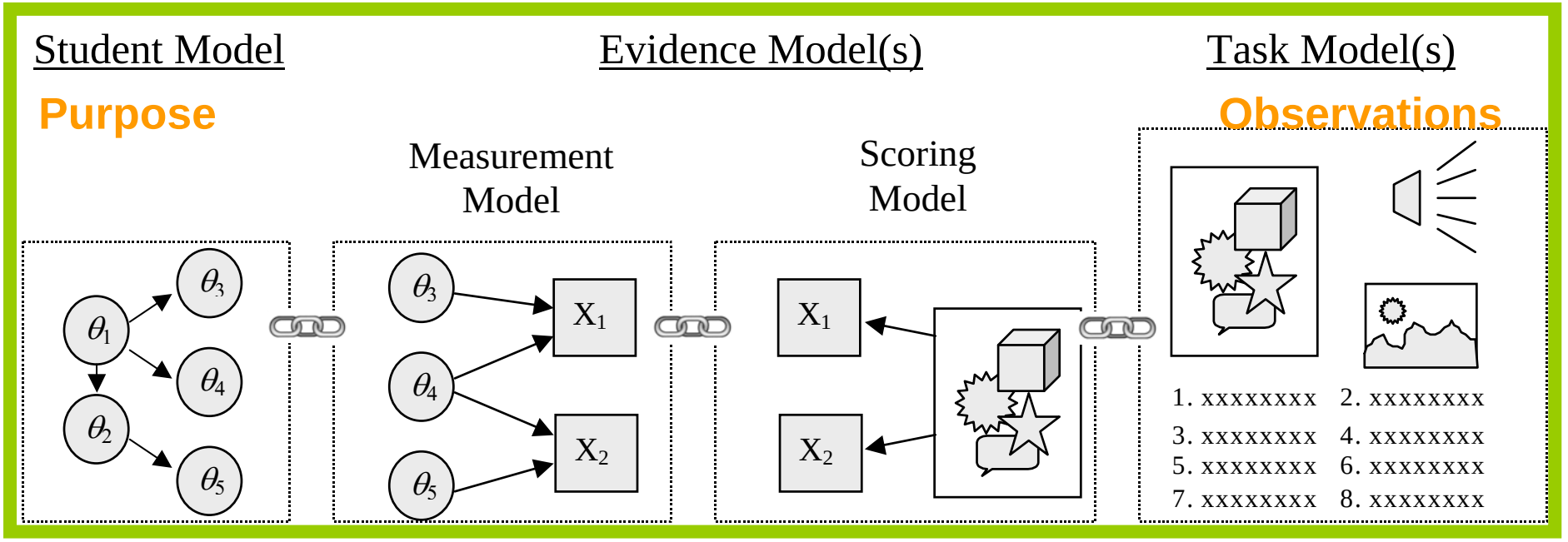
Evidence –Centered Assessment Design:
Using PADI
Session II:

Designing the Inferential Structures

Robert J. Mislevy
University of Maryland

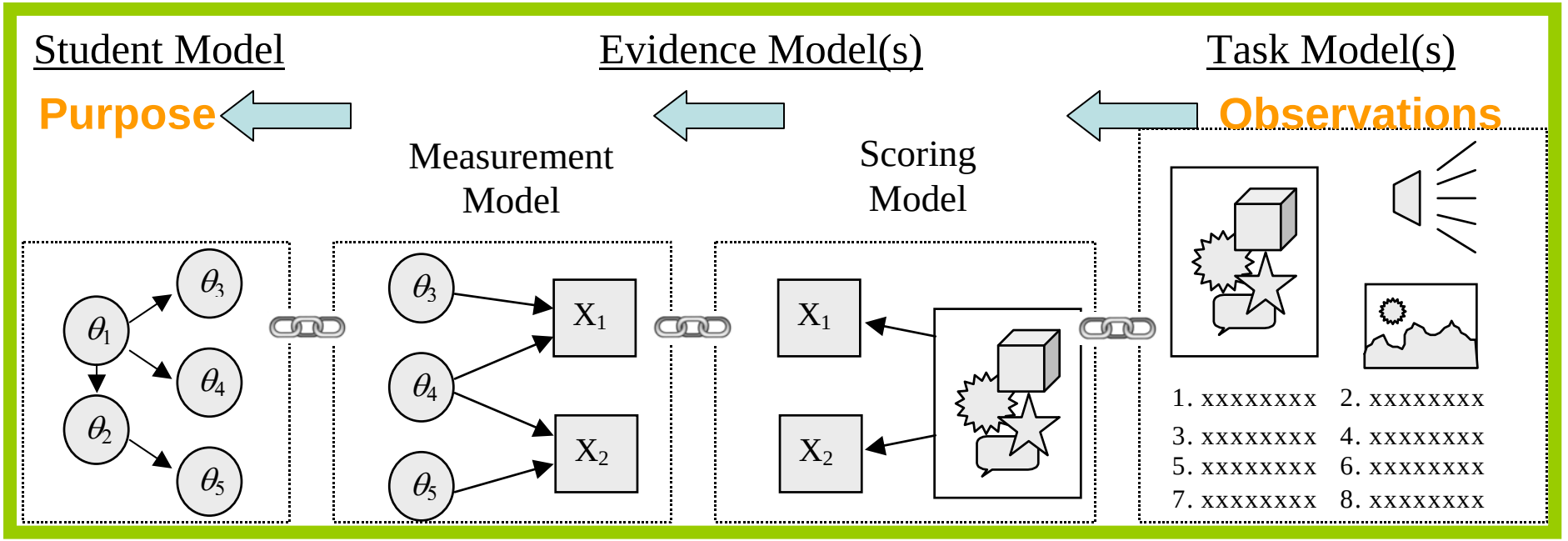
Geneva Haertel
Britte Cheng
Serena Villalba
SRI International

The Assessment Argument



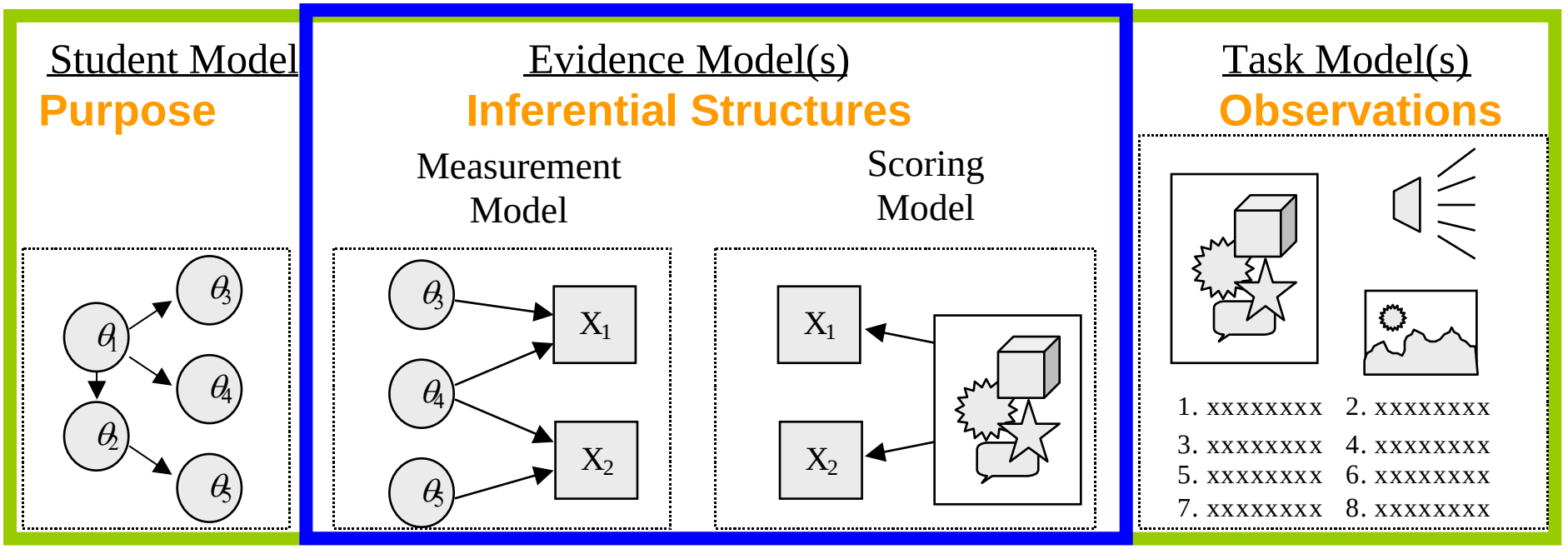
- 4 models to get right.

The Assessment Argument



- A chain is as strong as its weakest link.
- To draw inferences about student knowledge, start with the tasks.

Now: The Evidence Model



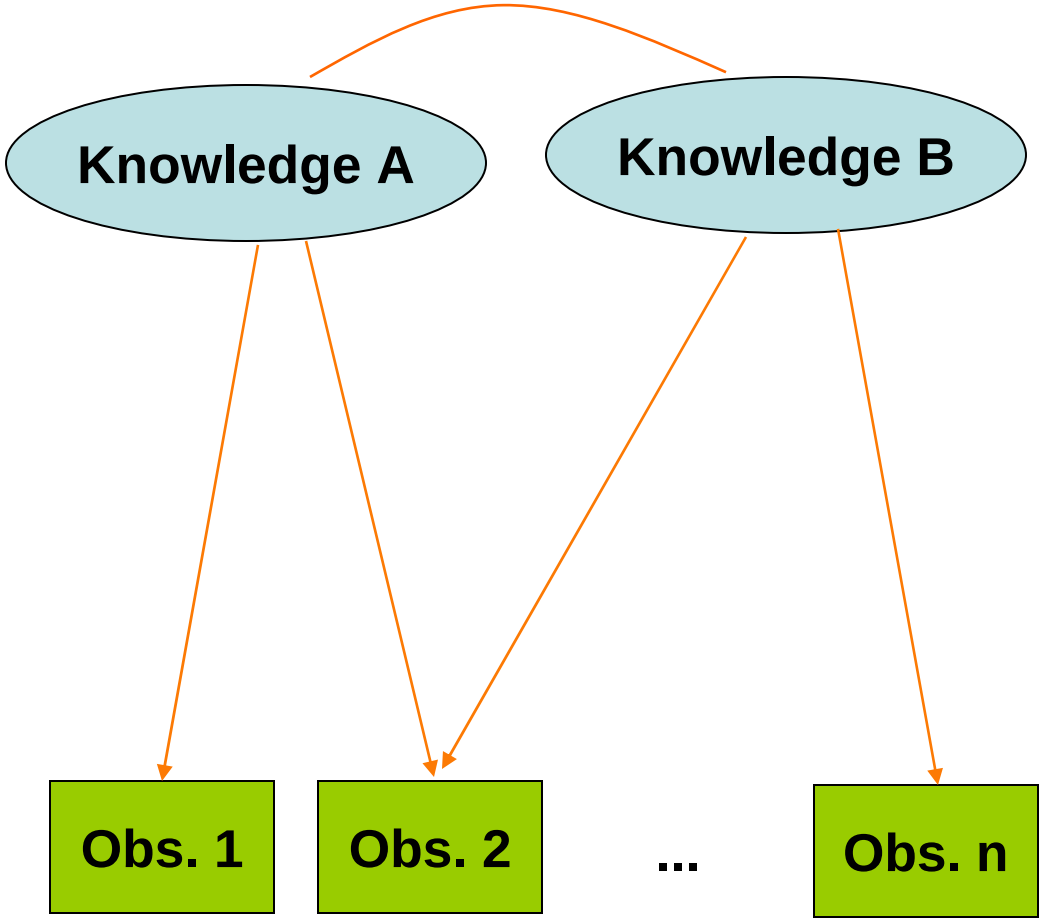
- Measurement: uses scores as evidence of the student model variables.
- Scoring: quantifies the observations to produce scores.

Underlying Theory

What is being assessed

"causes"

response evidence.



A simple inference…. dissected

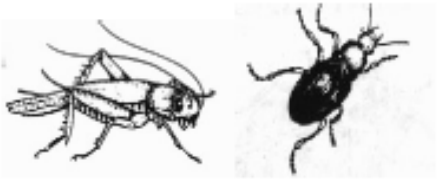
We want to know how well a student understands biodiversity content and science inquiry skills.

The assessment purpose.

1. Prompt to elicit evidence

Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below:

Group A:



Group B:

They want to place this fly  in either Group A or Group B. Where should this fly be placed?

A fly should be in **Group A /Group B**

Circle one

Name two physical characteristics that you used when you decided to place the fly in this group:

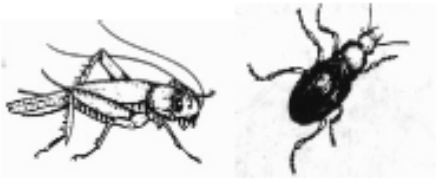
(a)

(b)

2. Gather the observations

Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below:

Group A:



Group B:

They want to place this fly  in either Group A or Group B. Where should this fly be placed?

A fly should be in **Group A**/**Group B**

Circle one

Name two physical characteristics that you used when you decided to place the fly in this group:

(a) spiders are bigger

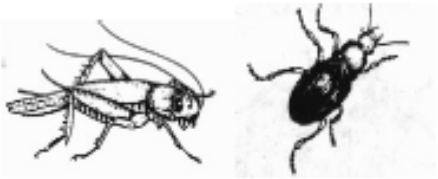
(b) insects have 6 legs

The observations.

3. Evaluate to accumulate evidence

Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below:

Group A:



Group B:

They want to place this fly  in either Group A or Group B. Where should this fly be placed?

A fly should be in **Group A**/**Group B**

Circle one

Name two physical characteristics that you used when you decided to place the fly in this group:

(a) spiders are bigger

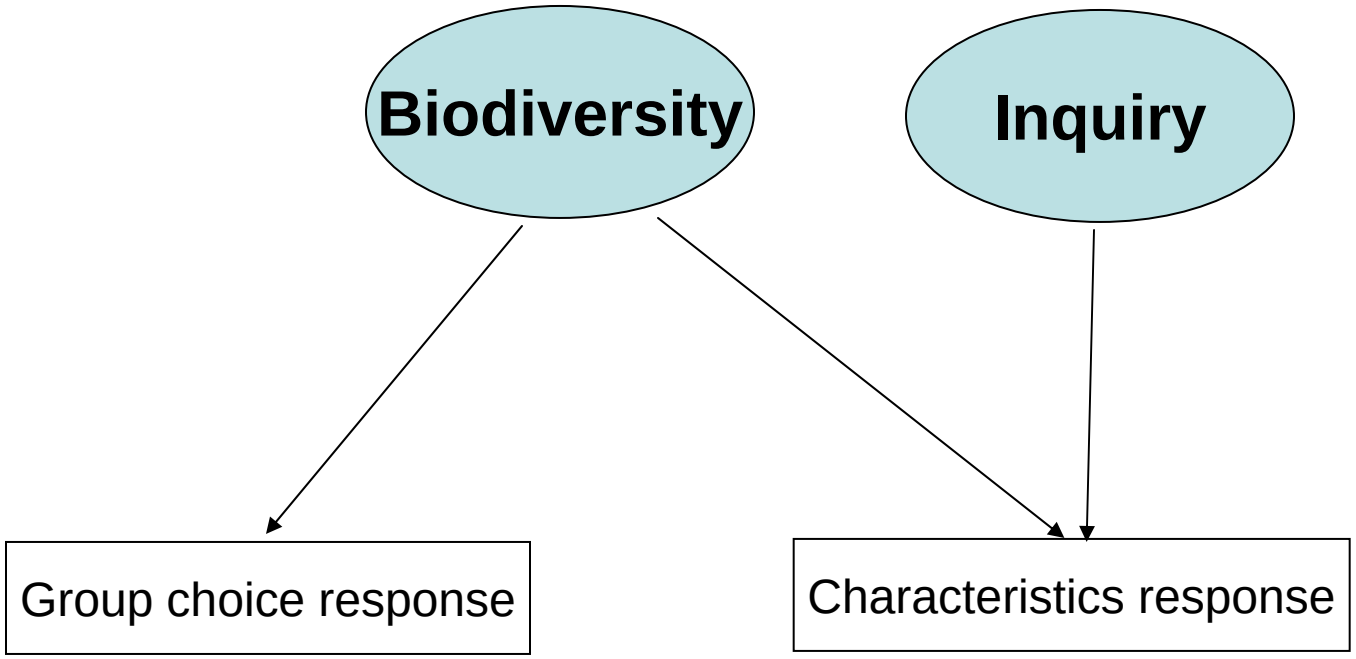
(b) insects have 6 legs

correct

The evidence.

partially correct

4. Draw inferences



The inferences.

Simple Evidence Model

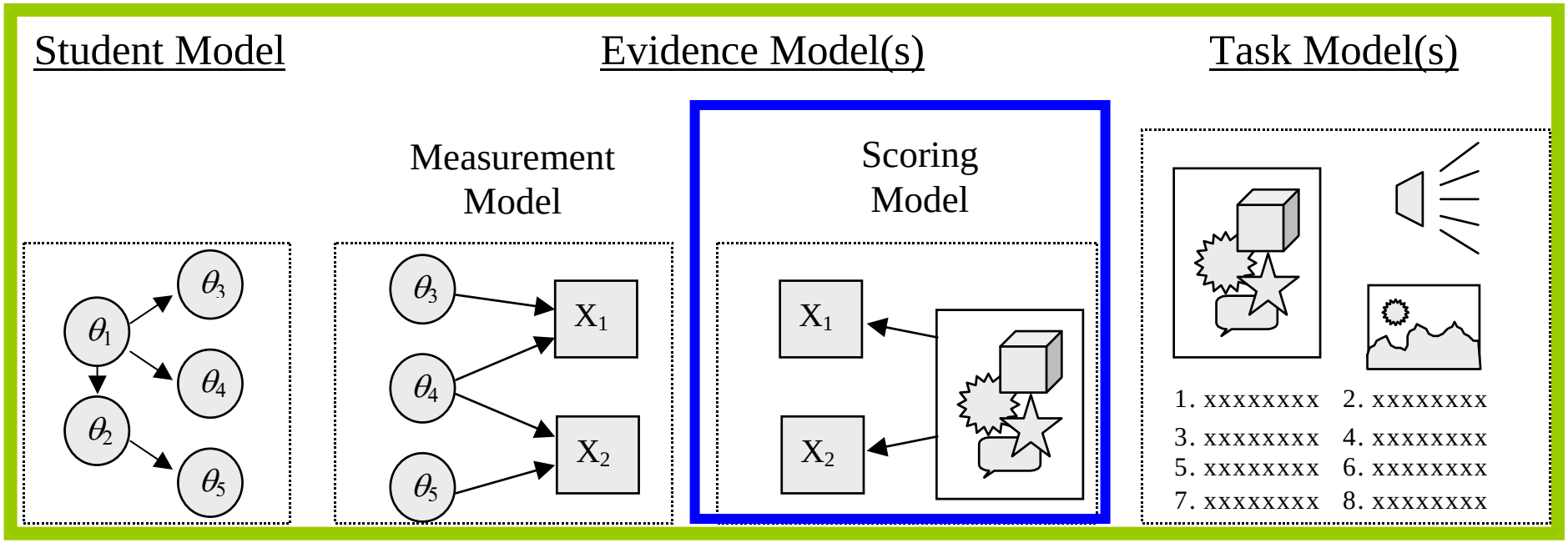
Student Model	Measurement Model	Scoring Model	Task Model
θ = science knowledge	$\hat{\theta} = \sum x$	1 = correct 0 = incorrect	Select correct response to queries requiring scientific knowledge.

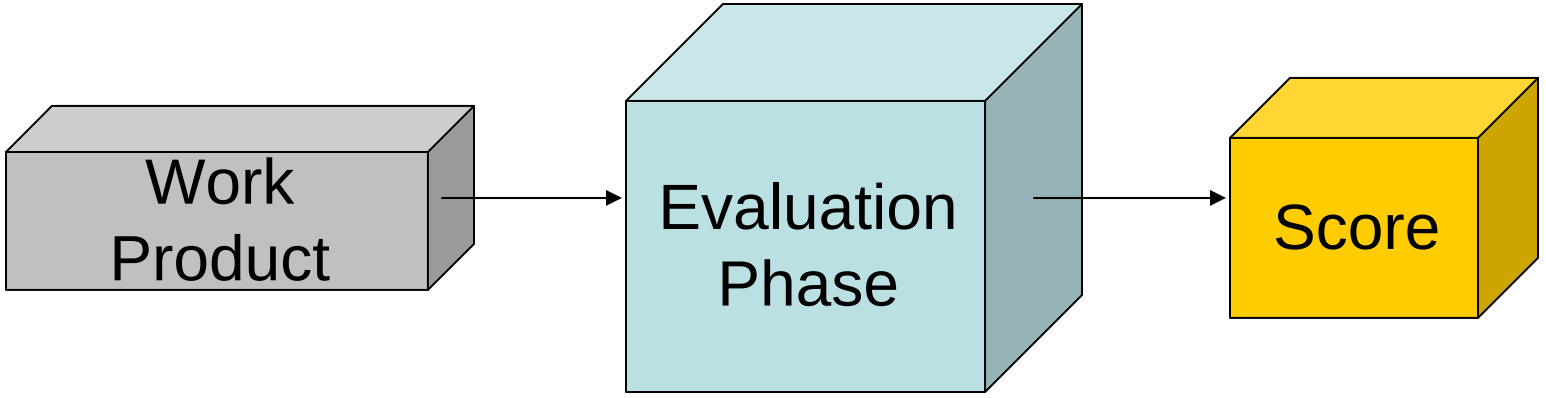
More Complicated Evidence Model

Student Model	Measurement Model	Scoring Model	Task Model
Multiple Correlated Constructs	Model response dependencies & multiple dimensions.	Integrate rater judgments, behavioral observation, artifacts, etc.	Simultaneously perform multiple, integrated tasks.

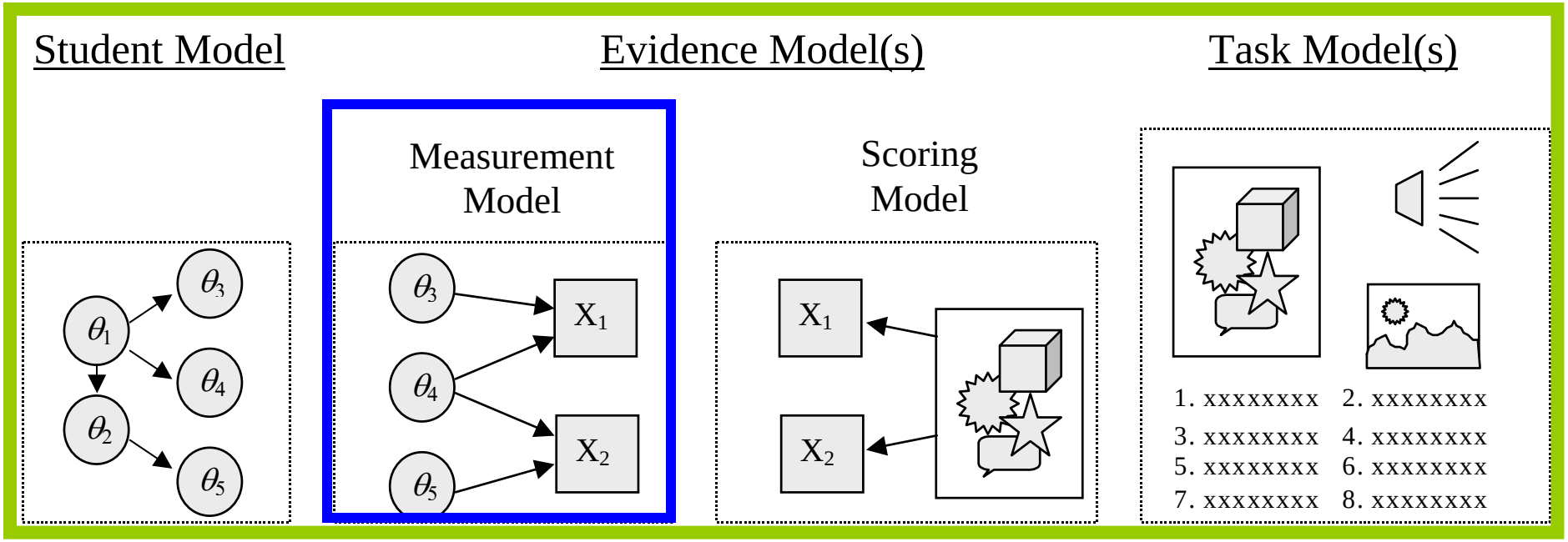
- Measurement & Scoring are the “glue” between students and tasks.

A Simple Scoring Model in PADI



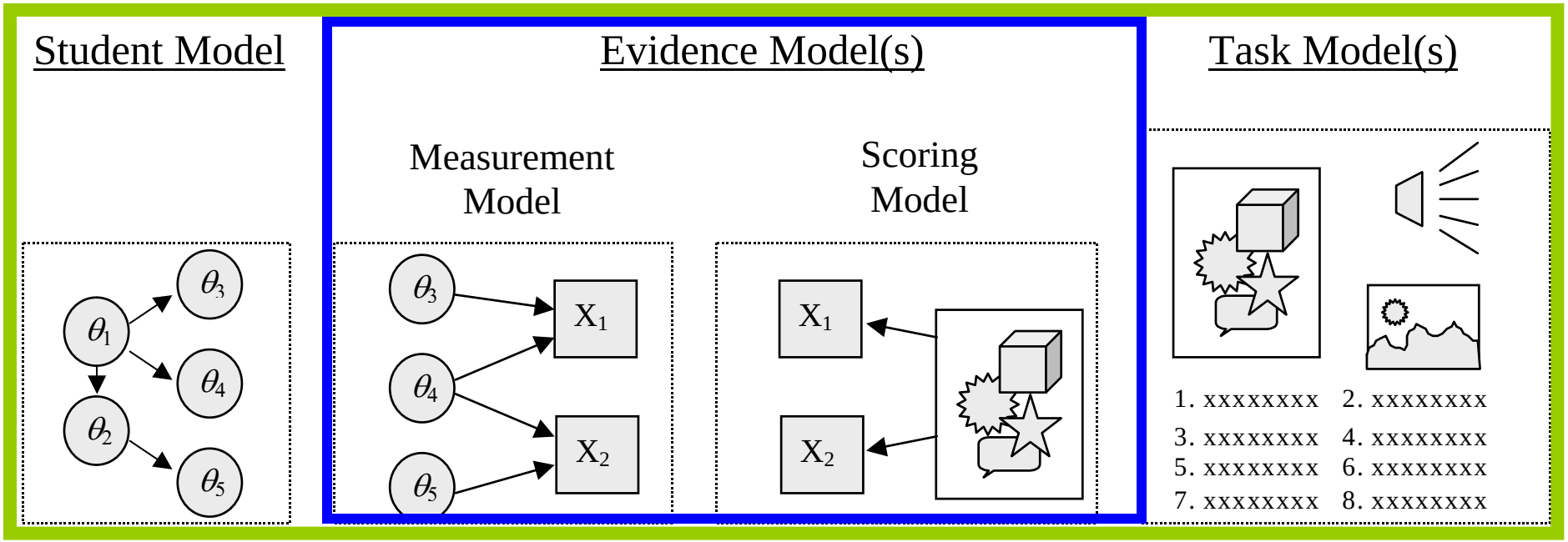


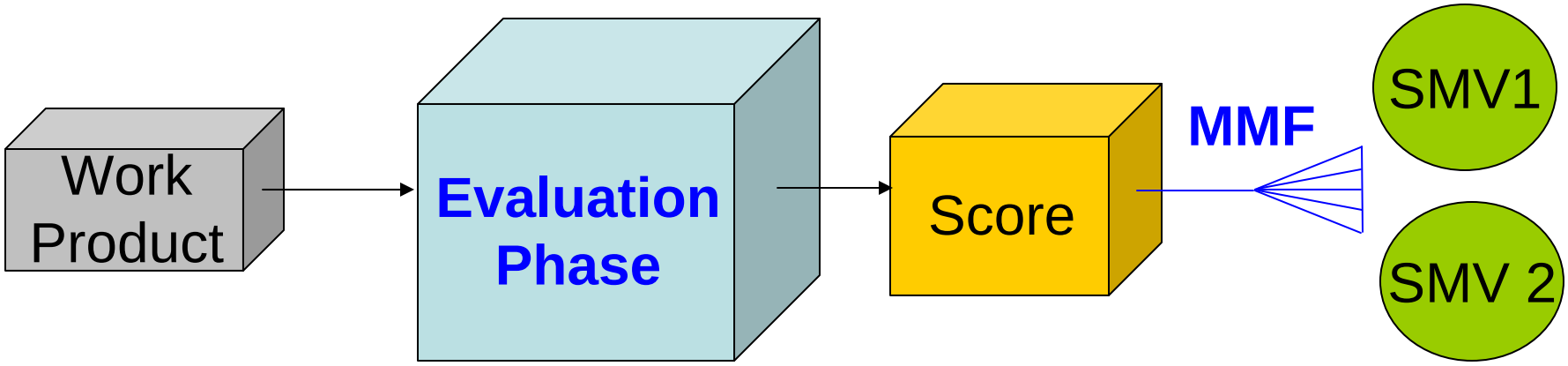
Simple Measurement Model in PADI



Score — Measurement Model Fragment — **Student Model Variable(s)**

An Evidence Model in PADI

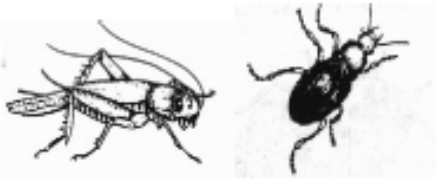




Revisit the BioKIDS Task

Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below:

Group A:



Group B:

They want to place this fly  in either Group A or Group B. Where should this fly be placed?

A fly should be in **Group A /Group B**

Circle one

Name two physical characteristics that you used when you decided to place the fly in this group:

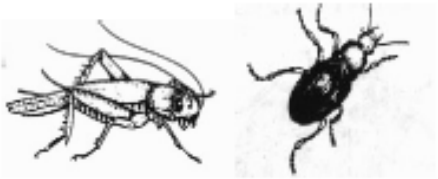
(a)

(b)

Revisit the BioKIDS Task (Work Products)

Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below:

Group A:



Group B:

They want to place this fly  in either Group A or Group B. Where should this fly be placed?

Claim

A fly should be in **Group A /Group B**

Circle one

Name two physical characteristics that you used when you decided to place the fly in this group:

(a)

(b)

Evidence

Evaluation Considerations

- Multiple, dependent tasks
- Single task may tap multiple dimensions of knowledge.
- A single task score can provide evidence of both dimensions and be independent of other scores.

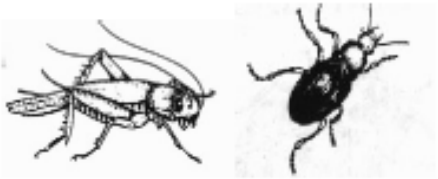
A Possible Evaluation Procedure

1. Score the Claim response
   - Correct (1) if selected Group A.
   - Incorrect (0) if selected Group B, multiple circles, or no response.
2. Score the Evidence response
   - Complete (2) if two correct responses with no incorrect responses.
   - Partial (1) if one correct response; or two correct responses *with additional incorrect responses.*
   - Incomplete (0) if other responses or no response.
3. Combine scores (to handle item dependence)

Revisit the BioKIDS Task (Student Model Vars)

Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below:

Group A:



Group B:

They want to place this fly  in either Group A or Group B. Where should this fly be placed?

A fly should be in **Group A /Group B**

Circle one

Biodiversity

Name two physical characteristics that you used when you decided to place the fly in this group:

(a)

(b)

Science Inquiry

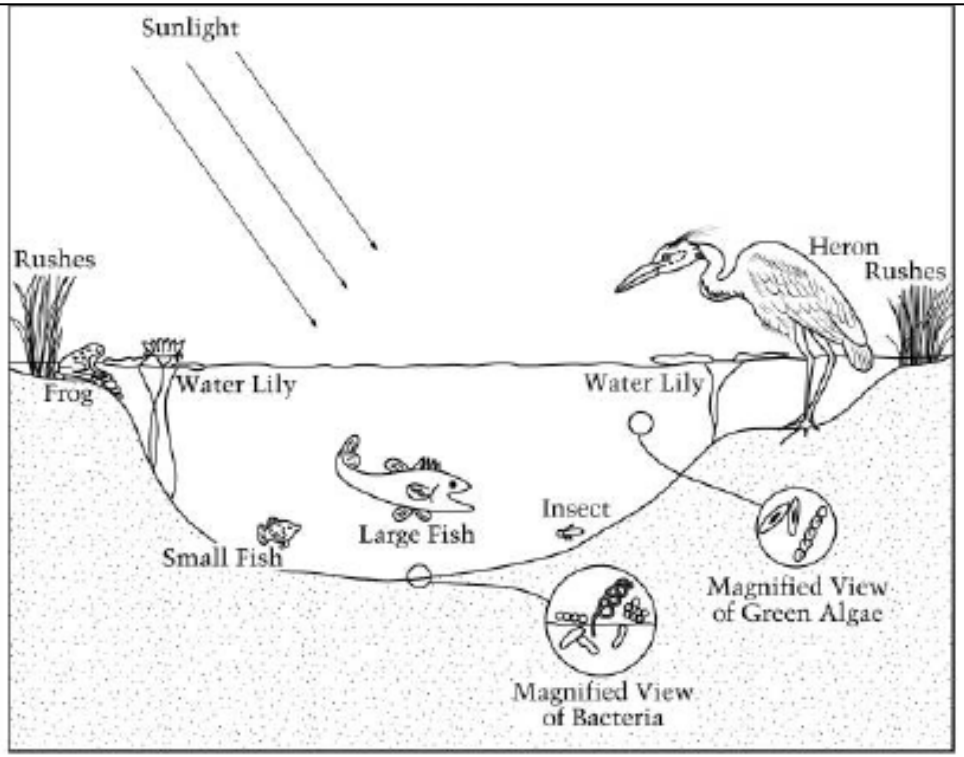
Measurement Model Considerations

- Associate Task Score(s) with Student Model Variables
  - Individual scores can be associated with different SMVs.
  - One score can be associated with multiple SMVs.

Possible Measurement Model Fragments

1. Two Task Scores
   - Claim score is an indicator of Biodiversity knowledge.
   - Evidence score is an indicator of Inquiry knowledge (no conditional correlation implied).
2. One Task Score
   - Combined score is an indicator of both Biodiversity and Explanations knowledge (conditional correlation possible).
   - Will return to this idea later

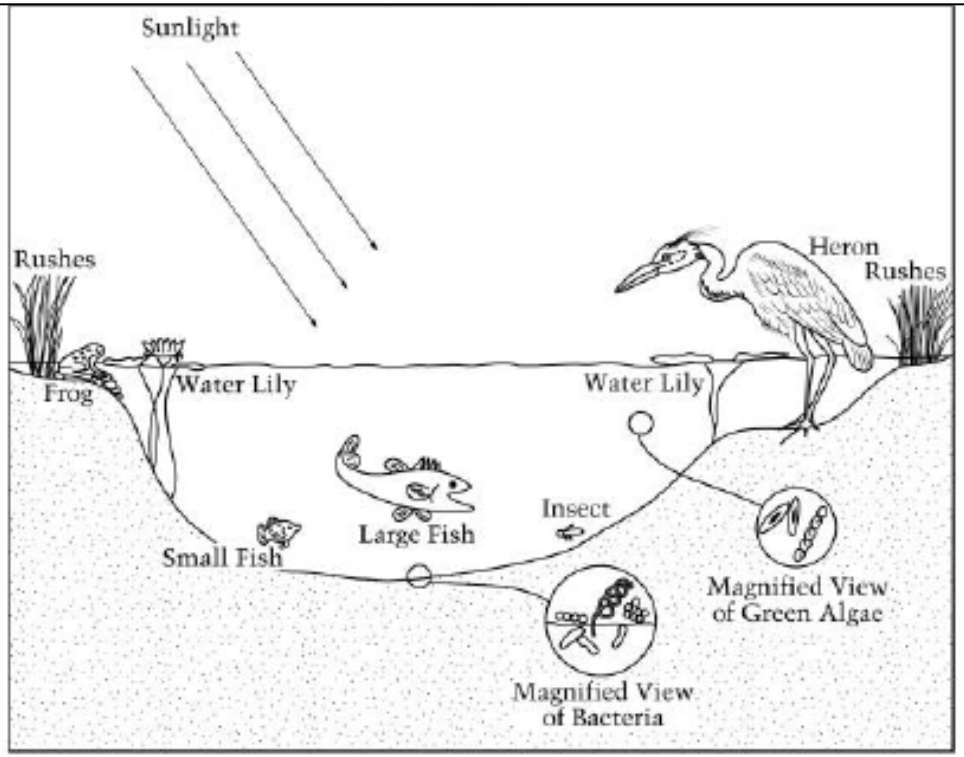
The assessment argument



http://nces.ed.gov/nationsreportcard/itmrls/

10. If all of the small fish in the pond system died one year from a disease that killed only the small fish, what would happen to the algae in the pond? Explain why you think so.

The Student Model



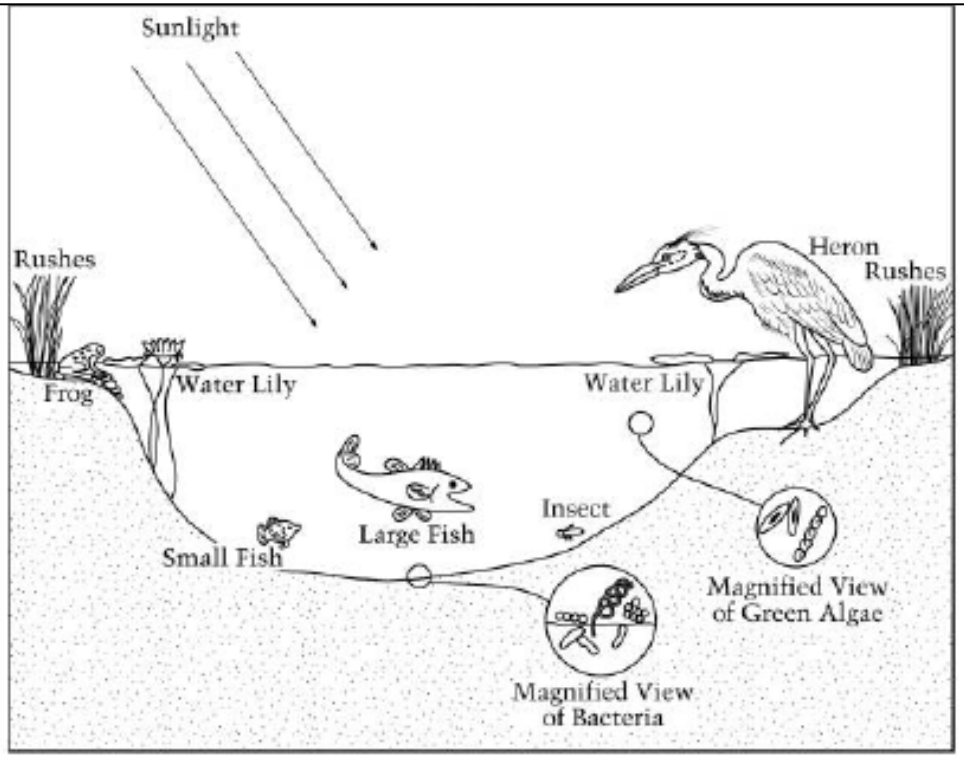
http://nces.ed.gov/nationsreportcard/itmrls/

10. If all of the small fish in the pond system died one year from a disease that killed only the small fish, what would happen to the algae in the pond? Explain why you think so.

Biodiversity

Inquiry

The Task Model



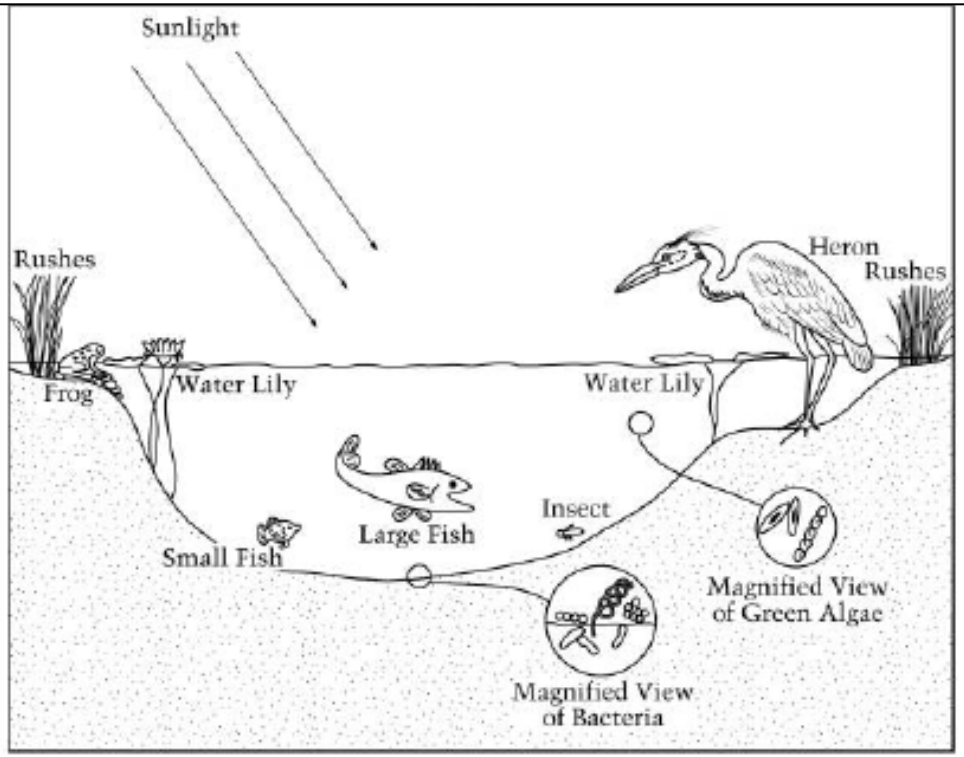
http://nces.ed.gov/nationsreportcard/itmrls/

10. If all of the small fish in the pond system died one year from a disease that killed only the small fish, what would happen to the algae in the pond? Explain why you think so.

Claim

Evidence

The Evidence Model - Scoring



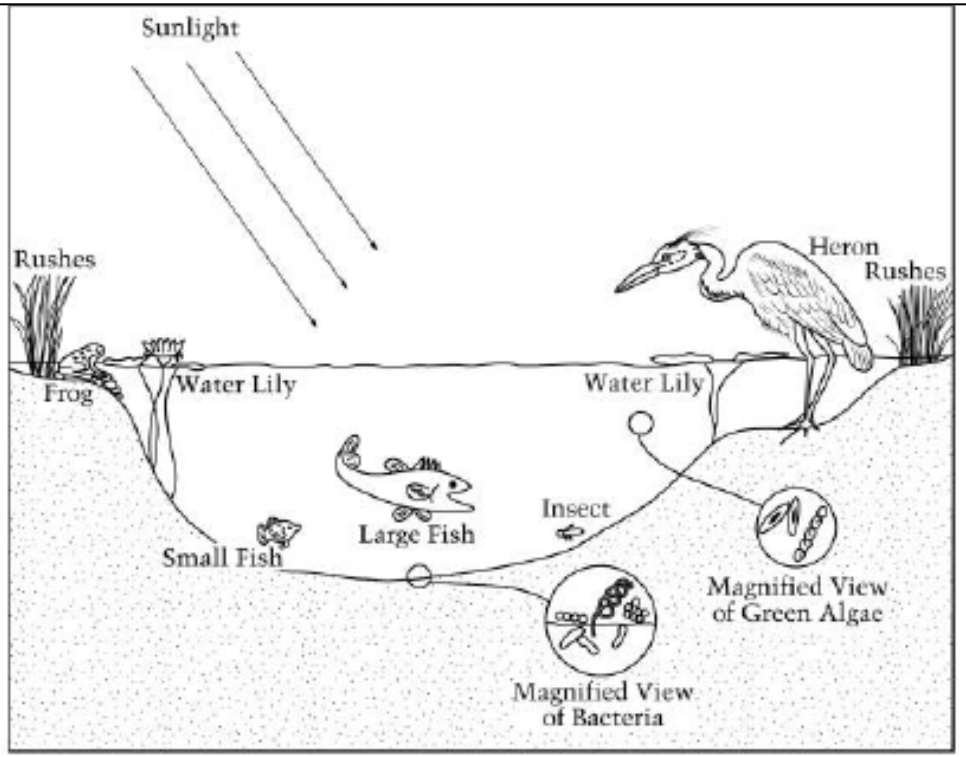
http://nces.ed.gov/nationsreportcard/itmrls/

10. If all of the small fish in the pond system died one year from a disease that killed only the small fish, what would happen to the algae in the pond? Explain why you think so.

Claim = 0,1

Evidence = 0, 1, 2

The Evidence Model - Measurement



http://nces.ed.gov/nationsreportcard/itmrls/

10. If all of the small fish in the pond system died one year fro disease that killed only the small fish, what would happen to the algae in the pond? Explain why you think so.

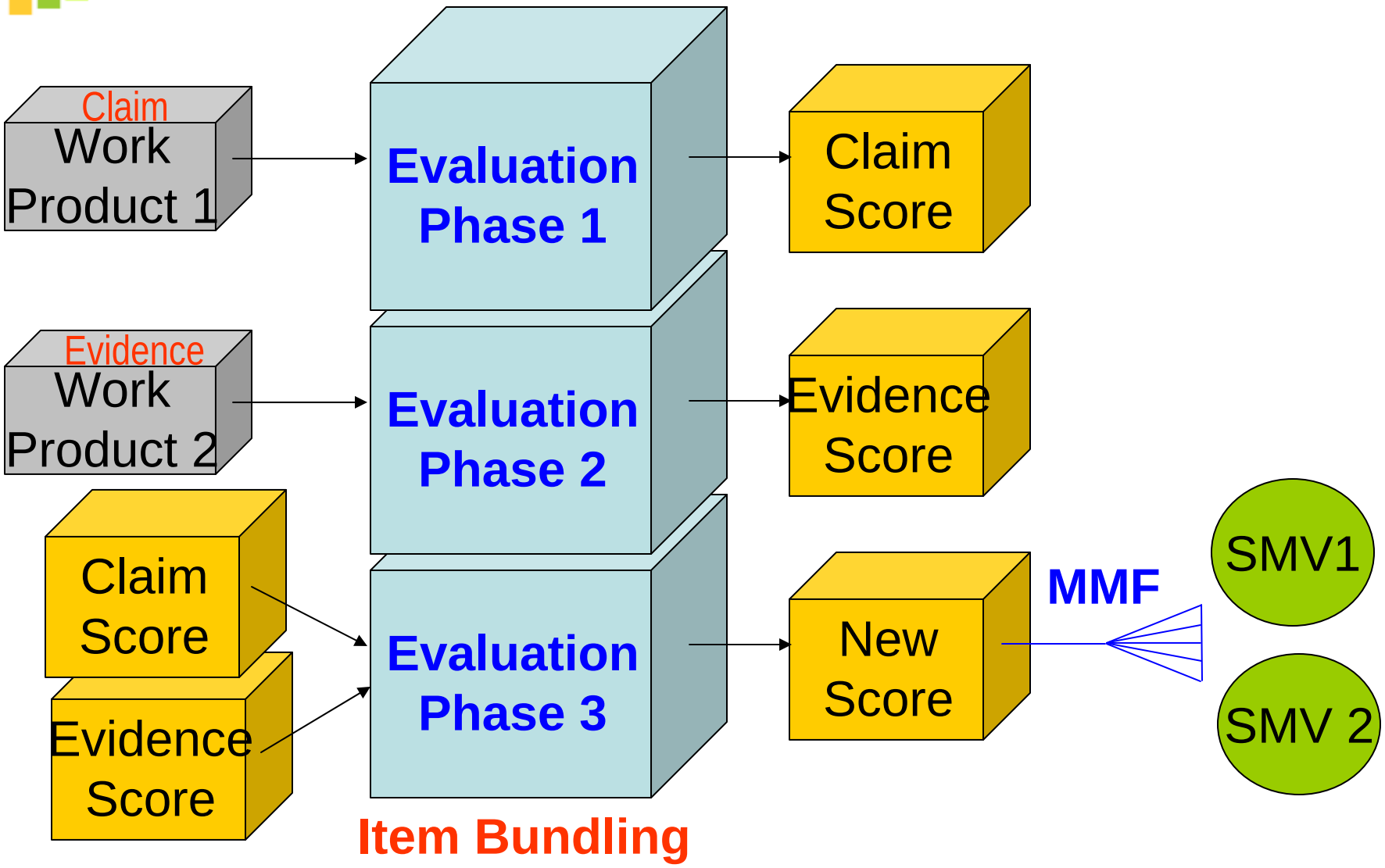
Claim score indicates Biodiversity

Evidence score indicates Inquiry

Possible Measurement Model Fragments

1. Two Task Scores
   - Claim score is an indicator of Biodiversity knowledge.
   - Evidence score is an indicator of Inquiry knowledge (no conditional correlation implied).
2. One Task Score
   - Combined score is an indicator of both Biodiversity and Explanations knowledge (conditional correlation possible).
   - Will return to this idea later

A More Complex Evaluation Procedure



Item Bundling Evaluation Phase

- Decide how to combine scores.

Claim	Evidence	Combined	Biodiversity	Inquiry
0	anything	0	0	0
1	0	1	1	0
1	1	2	1	1
1	2	3	1	2

- Designer specifies loadings via a Scoring Matrix within the Measurement Model Fragment.

Evidence Models – Take Away Points

- The ability to assign a score to a performance does not an evidence model make.
- Estimating IRT parameters does not an evidence model make.
- 4 models to align: lots of bookkeeping
- In principle, we have an audit trail (a thread of an *assessment argument*) from any task behavior to SMVs.

Floating Pencil Task Outline

After 0%

After 25%

From All Solutions

1. **Explain** why the pencil floats in 0% solution

2. **Compare** results for 0 % and 25% salt solution. (select statement)

3. **Explain** why the pencil floats at different levels in 0% and 25% solution.

4. **Predict** what will happen to the pencil if more salt were added to the solution. (select statement)

5. **Measure** the height of the pencil floating in 0%, 25%, & X% salt solution.

6. **Explain** why it is important to take two measurements for each solution.

7. **Average** two measurements for each solution.

8. **Graph** the average pencil height against salt concentration for 0% and 25% solution.

9. **Describe** the relation-ship shown in graph (select statement).

10. **Interpolate** to find the concentration of the unknown solution

Follow instructions to float a pencil in 0%, 25% & X% salt solution

Floating Pencil: Try it out

- Your Task:
  - Select 1 or 2 discrete tasks in FP
  - Select a schema of SMVs
  - Define rules to quantify observational performance on tasks (scoring model)
  - Link scores to student model variables (measurement model)
  - Share with group discussion

One Possible Mapping