

Reconceptualizing Items: From Clones and Automatic Item Generation to Task Model Families

Richard M. Luecht, Ph.D., UNC-G

Matthew J. Burke, Ph.D., ABIM

Outline of the Talk

- Rationale
- Reconceptualizing “items” and test content
- Item models and automatic item generation (AIG): mechanisms for mass producing items
- Cognitive task models and “item families”: an engineering approach to scale and test development
- Quality control (QC) for item families

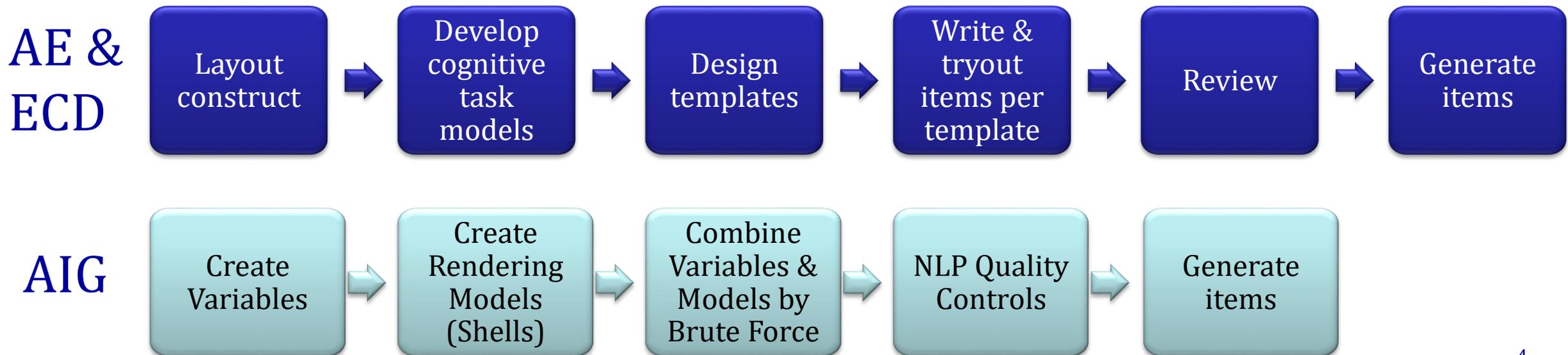
Why Use Automated Item Generation and Principled Item Design Technologies?

“The demand for large numbers of items is challenging to satisfy because the traditional approach to test development uses the item as the fundamental unit of currency. That is, each item is individually hand-crafted—written, reviewed, revised, edited, entered into a computer, and calibrated—as if no other like it had ever been created before.”

Dragow, Luecht & Bennett, *Educational Measurement*, 4th Edition, p. 473

So...Item and Test Design are Changing

- Traditionally, the quality of test item generation has been dependent on the experience and interpretation of content specifications by item writers (Schmeiser & Welch, 2006)
- Principled item design (Bennett, 2001; Irvine, 2002) is rapidly evolving from theory to practical implementation



The Evolution of “Items”

Item Writing/Editing as a *Craft*?



Item Shells

```

<Patient.article><Patient.description.age>
<Patient.description.occupation>
"comes to" <Setting.description> "complaining of"
<Patient.ailment.symptom1> <Patient.ailment.symptom1.duration>
<Patient.ailment.symptom2> <Patient.ailment.symptom2.duration>
<Patient.history.activity.recent>
<Patient.physicalexam.temp=# C, (convert(C,F))>
<Patient.physicalexam.pulse=#/min>
<Patient.physicalexam.respiration=#/min>
<Patient.physicalexam.bp=#1/#2>
<Patient.physicalexam.symptom1>
<Patient.physicalexam.symptom2> "What is the most likely cause of"
<Patient.ailment.prime_symptom> "?"
    
```

Different Item

Generation

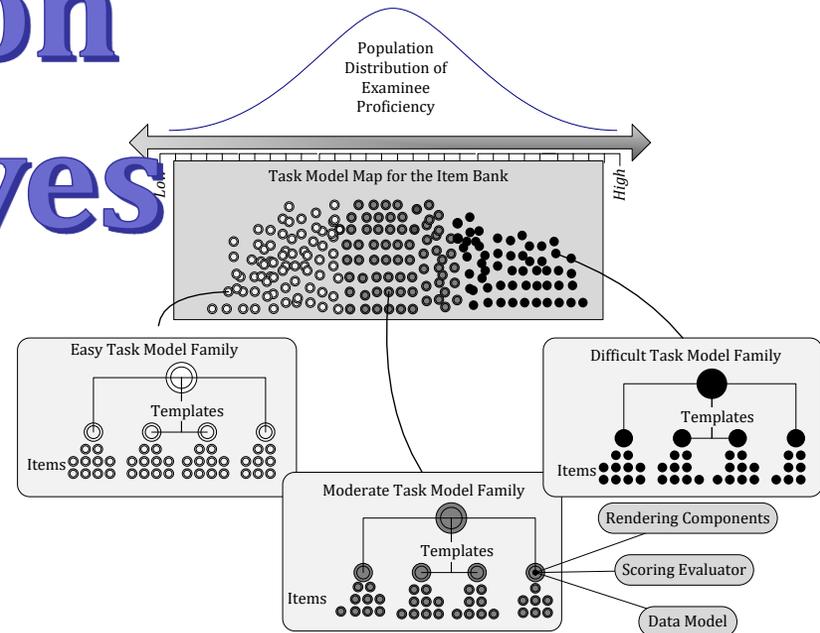
Perspectives

Task Model Families

N-Layered AIG & Mass Item Production Models

Solve (low complexity function | several number of operations. explicit application. three unique variable)

| Item Model Variables | |
|----------------------|---|
| Stem | Simplify the following in equation: $\text{Log}(10^{112} + 10^{73} - 10^{24})$ |
| Elements | I1 Value Range: 2 - 5 by 1 |
| | I2 Value Range: 2 - 3 by 1 |
| | I3 Value Range: 2 - 5 by 1 |
| | I4 Value Range: 2 - 5 by 1 |



Questions 10 – 12

Choose the appropriate letters A, B, C or D.

Write your answers in boxes 10-12 on your answer sheet.

10 Research completed in 1982 found that in the United States

- A reduced the productivity of farmland by 20 per cent
- B was almost as severe as in India and China
- C was causing significant damage to 20 per cent of farmland
- D could be reduced by converting cultivated land to pasture

11 By the mid-1980s, farmers in Denmark

- A used 50 per cent less fertiliser than Dutch farmers
- B used twice as much fertiliser as they had used in 1950
- C applied fertiliser much more frequently than in 1950
- D more than doubled the amount of pesticide used

12 Which one of the following increased in New Zealand after 1980?

- A farm incomes
- B use of fertiliser
- C over-stocking
- D farm diversification

Click on each of the five balls to toss them on to the skeeball board. Observe how many points you receive with each toss.

Calculate the following values. Type your answers in the boxes below.

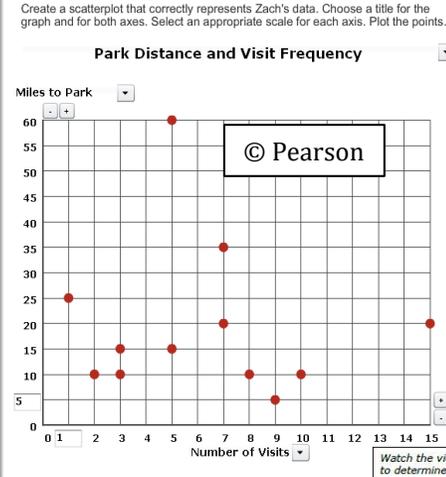
The Mean:

The Mode:

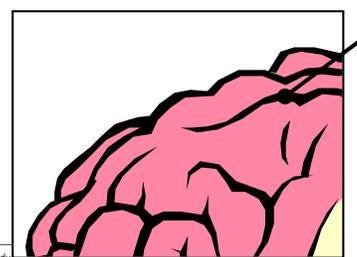
The Median:

© Pearson

| Number of Visits per Year | Distance to Park (miles) |
|---------------------------|--------------------------|
| 1 | 25 |
| 2 | 10 |
| 7 | 35 |
| 3 | 10 |
| 5 | 60 |
| 7 | 20 |
| 5 | 15 |
| 15 | 20 |
| 8 | 10 |
| 10 | 10 |
| 3 | 15 |
| 9 | 5 |




Where is the in the cerebral cortex's parietal lobe? (Click on the appropriate area of the picture, below)



Parietal Lobe

You are working for a spy agency and must design a capsule that can detect Molecule X in a solution. The capsule will have a semipermeable membrane and will be filled with a solution.

You are given the following information:

- Molecules X and Y are both in a solution, and will turn blue when present along with X and Y. It will neutralize the reaction and the solution will not turn blue.
- Molecules X, Y, and Z all dissolve in water.

You must find the right membrane, capsule solution, and beaker solution. A laboratory is provided with three different capsule membranes and samples of Molecules X, Y, and Z. You can perform experiments to determine the properties of the capsule membranes and the molecules.

Select a capsule membrane, capsule solution, and beaker solution. The capsule will be submerged in the beaker solution, and the results will be recorded in the table.

Choose Capsule Membrane

- Membrane A
- Membrane B
- Membrane C

Choose Capsule Solution

- 10% Molecule X
- 10% Molecule Y
- 10% Molecule Z

Choose Beaker Solution

- 10% Molecule X
- 10% Molecule Y
- 10% Molecule Z

©American Institutes of Research

Watch the video of an actual roller coaster ride and the animation showing the same roller coaster. You will be asked to determine at what point the roller coaster car has the most kinetic energy.

Select the activity you would like to do. Then use controls below the images to view the video or the animation:

Video Animation Answer Space

At what point does the roller coaster car have the greatest amount of kinetic energy?

To answer, select "Answer Space." Then click on the red roller coaster track in the answer space to show the correct location. To change your answer, click on a different point along the red roller coaster track.

Look carefully at the Utah ecosystem shown. Sort the organisms in the ecosystem into three groups: producers, consumers, and decomposers.

There are two ways to explore the scene.

1. Move the mouse cursor over the scene to view organisms more closely.
2. Hover the mouse cursor over the name of an organism in the Word Bank to highlight the organism in the ecosystem.

Click and drag the names of organisms from the Word Bank to the correct places in the chart below.

- You may move the words in the chart after you have placed them.
- To complete the question, place all the organisms in the chart.

Word Bank

- Bacteria
- Cottonwood
- Fungi
- Grasses
- Jackrabbit
- Mule deer
- Red-tailed hawk
- Sagebrush

| Producers | Consumers | Decomposers |
|-----------|-----------|-------------|
| | | |

Clear all

amusement park offers Ferris wheel rides to 10 and adults. A rider is shown on the Ferris wheel.

Write a function $h(t)$ that models the height, h , of the rider at any point on the Ferris wheel in terms of t , time in minutes.

Click the buttons to create your answer.

©American Institutes of Research

A junior staff accountant in your department seems to be confused about what constitutes an accounting change for financial reporting purposes. In a brief memorandum to the junior accountant, describe the three basic types of accounting changes. Type your description in the space below using the word processor provided.

REMINDER: Your response to this prompt will be graded for both technical content and writing skills. For writing skills, you should demonstrate the ability to develop your ideas, organize them, and express them clearly. Do not convey information in the form of a table, bullet point list, or other abbreviated presentation.

©AICPA

Initial History

Reason for Visit: Chest pain, respiratory distress

History of Present Illness: The patient, a 45-year-old accountant, is brought to the emergency department by ambulance from the trucking company where he works. Oxygen was administered during transport. About 10 minutes before arrival, he developed excruciating, sharp pain in the right side of his chest that radiated to his left arm and jaw. The pain increased with respiration. He is unable to answer further questions. A family member who accompanied the patient to the hospital says that this never happened before. The patient has had hypertension and asthma for years.

All other history is obtainable.

You are provided an Initial History at the start of the case. The history is complete, assume any missing information is contributory. When you have finished reading the initial history, select OK to continue.

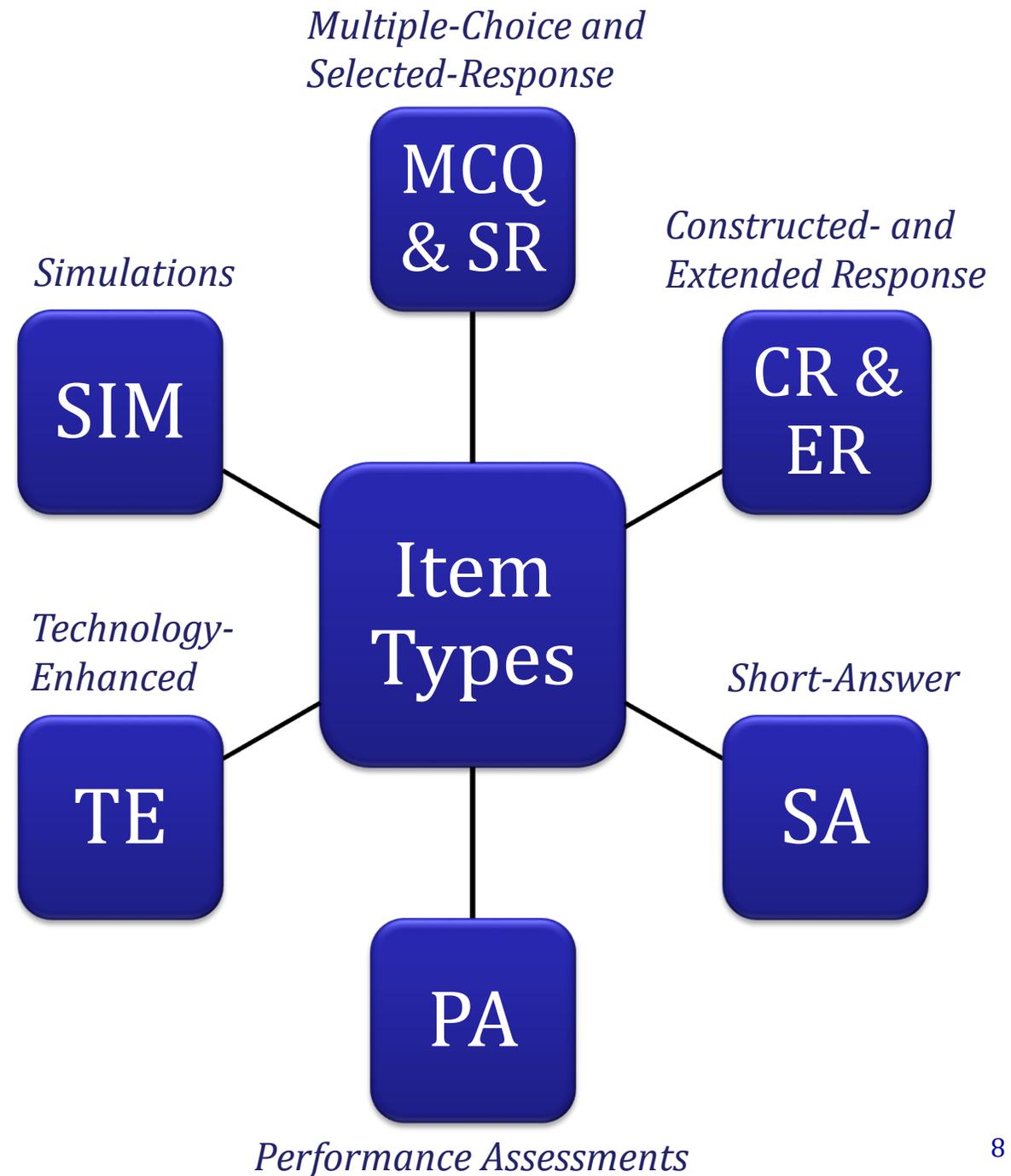
To continue, press the space bar or Enter key.

©NBME

What is an "item"?

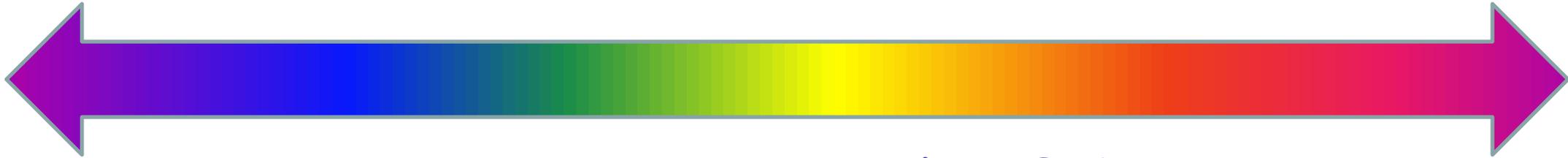
Complications of Item Types

- Stimuli, prompts and problem instructions
- Exhibits
- Auxiliary tools/resources
- Response capturing
- Response data
- Scoring evaluators



Unconstrained

Constrained



Actual Performance

Work Samples

Interactive Simulations

Essays, PBAs, Free-Response (FR) and CR Items

Short-Answer (SA) Items

Technology-Enhanced (TE) Items

Long-Option List Selected-Response Items (Incl. "Pick-N's")

Multiple-Choice (MC) Items

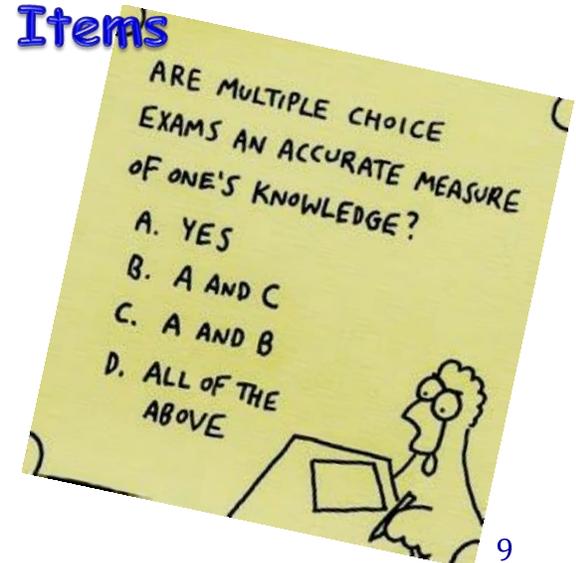
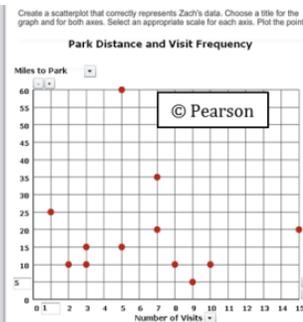
True-False or Binary Choice Items



Zach surveyed a group of people visiting state parks to determine if the distance they lived from the park affected how frequently they visited the park.

Zach's data is shown in the table below.

| Number of Visits per Year | Distance to Park (miles) |
|---------------------------|--------------------------|
| 1 | 25 |
| 2 | 10 |
| 7 | 35 |
| 3 | 10 |
| 5 | 60 |
| 7 | 20 |
| 5 | 15 |
| 15 | 20 |
| 8 | 10 |
| 10 | 10 |
| 3 | 15 |
| 9 | 5 |



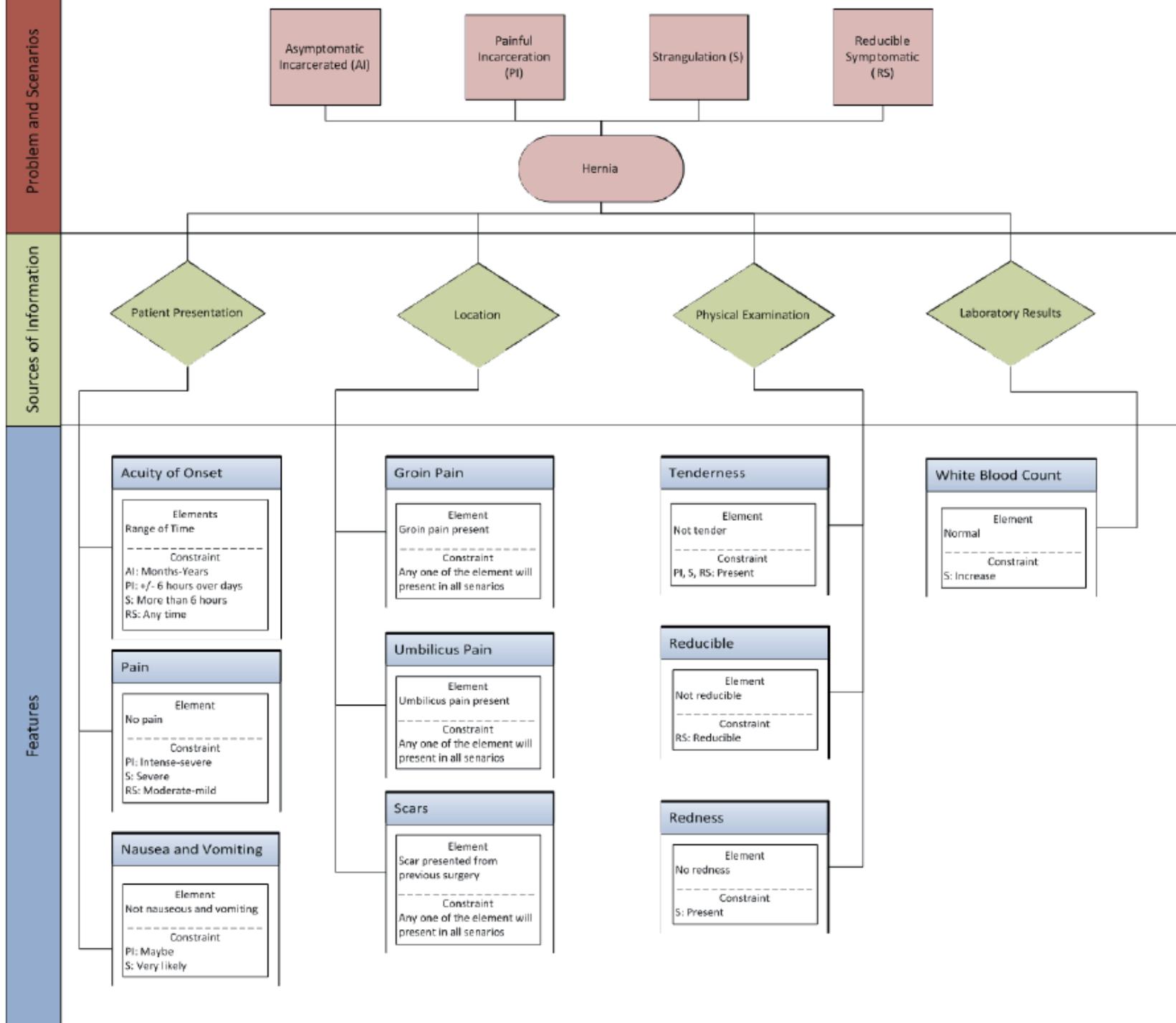
Automatic Item Generation (AIG) for Enhanced Multiple- Choice Item Production

AIG in Three Steps*

- The content required for the generated items is identified by test development specialists and defined as a *cognitive model*
- An *item model* is developed by the test development specialists to specify where content is placed in each generated item
- In Step #3, computer-based algorithms are used to place the content specified in Step #1 into the item model developed in Step #2

* Gierl, Lai & Turner (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, 46,757-765

Step #1. Documenting the Item Content



Step #2. Generating an Item Model

A 25-year-old man presented with a mass in the left groin. It occurred suddenly 2 hours ago while lifting a piano. On examination, the mass is firm and located in the left groin and lab work came back with normal results. Which of the following is the next best step?

A [**AGE**]-year-old [**GENDER**] presented with a mass [**PAIN**] in [**LOCATION**]. It occurred [**ACUITYOFONSET**]. On examination, the mass is [**PHYSICALFINDINGS**] and lab work came back with [**WBC**]. Which of the following is the next best step?

[**AGE**] (Integer): From 25.0 to 60.0, by 5.0

[**GENDER**] (String): 1: man 2: woman

[**PAIN**] (String): 1: 2: and intense pain 3: and severe pain 4: and mild pain

[**LOCATION**] (String): 1: the left groin 2: right groin 3: the umbilicus 4: an area near a recent surgery

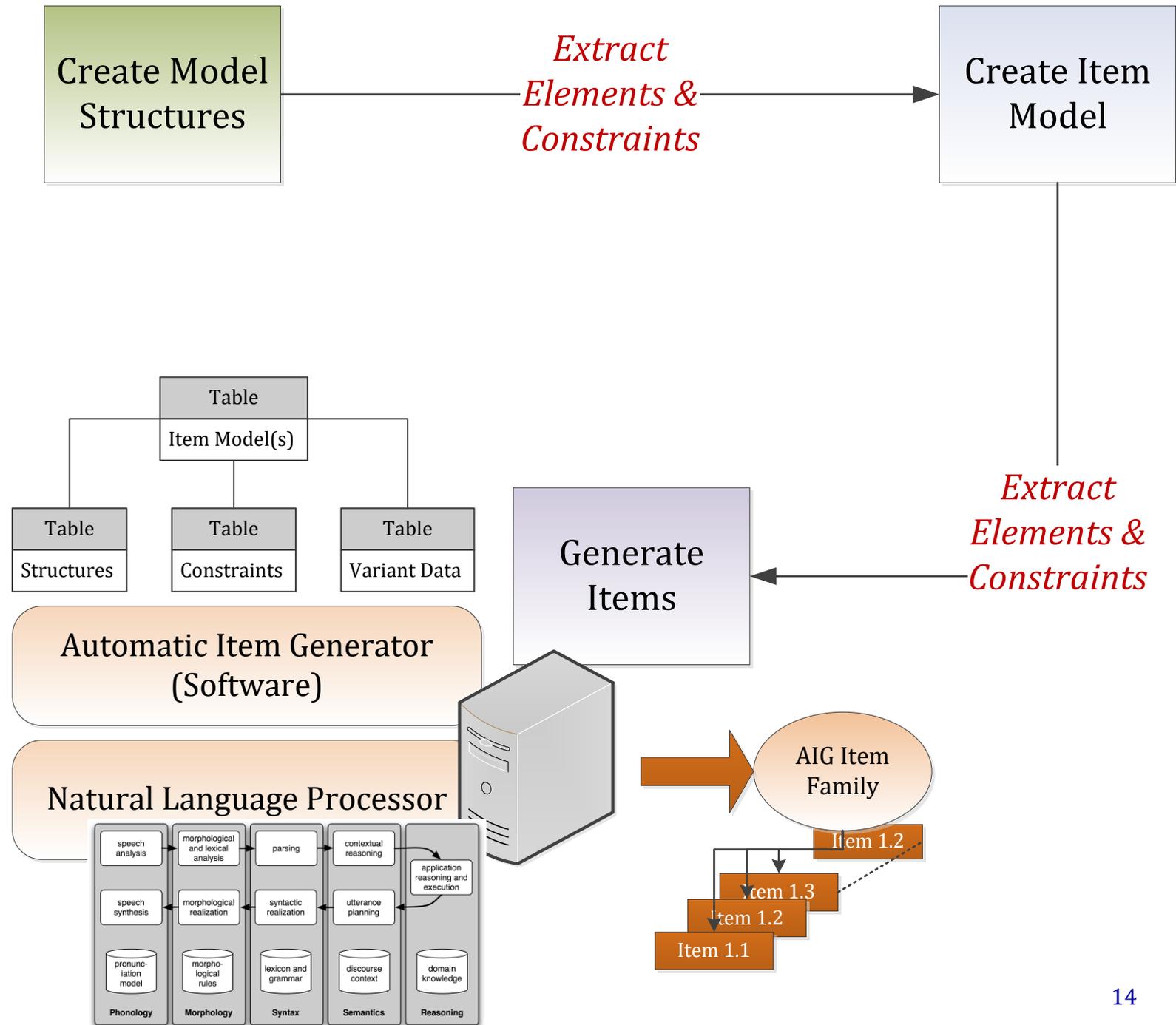
[**ACUITYOFONSET**] (String): 1: a few months ago 2: a few hours ago 3: a few days ago 4: a few days ago after moving a piano

[**PHYSICALFINDINGS**] (String): 1: protruding but with no pain 2: tender 3: tender and exhibiting redness 4: tender and reducible

[**WBC**] (String): 1: normal results 2: normal results 3: elevated white blood cell count 4: normal results

Contextual features: exploratory surgery; reduction of mass; hernia repair; ice applied to mass

Step #3: Submitting Template(s), Elements and Constraints to an *Item Generator*



Sample AE Math Task Model and Templates

Solve (low complexity function | several number of operations. explicit application. three unique variable)

| | Item Model Variables |
|-----------------|--|
| <i>Stem</i> | Simplify the following in equation: $\text{Log}(10^{I1^{I2}} + 10^{I3} - 10^{I4})$ |
| <i>Elements</i> | I1 Value Range: 2 – 5 by 1 I2 Value Range: 2 – 3 by 1 I3 Value Range: 2 – 5 by 1 I4 Value Range: 2 – 5 by 1 |

Task Model

Item Template

Lai, H.; Gierl, M. & Alves, C. (2010). *Generating Items under the AE Framework*. Invited symposium paper at the Annual Meeting of NCME, Denver

AE Item Production

| Item Template # | Range of Elements | | | | | Number of items |
|-----------------|-------------------|-----------|-----------|-----------|-----------|-----------------|
| | Element 1 | Element 2 | Element 3 | Element 4 | Element 5 | |
| Mathematics | | | | | | |
| 1 | 5 | 8 | 9 | | | 355 |
| 2 | 2 | 3 | 3 | | | 18 |
| 3 | 8 | 6 | 4 | 4 | | 762 |
| 4 | 20 | 9 | 33 | | | 5940 |
| 5 | 3 | 3 | 3 | | | 27 |
| 6 | 3 | 2 | 2 | | | 12 |
| 7 | 8 | | | | | 8 |
| 8 | 4 | 4 | 4 | 4 | 4 | 896 |
| 9 | 5 | 3 | 13 | 11 | | 2132 |
| 10 | 4 | 2 | 4 | 4 | | 128 |

Lai, H.; Gierl, M. & Alves, C. (2010). *Generating Items under the AE Framework*. Invited symposium paper at the Annual Meeting of NCME, Denver

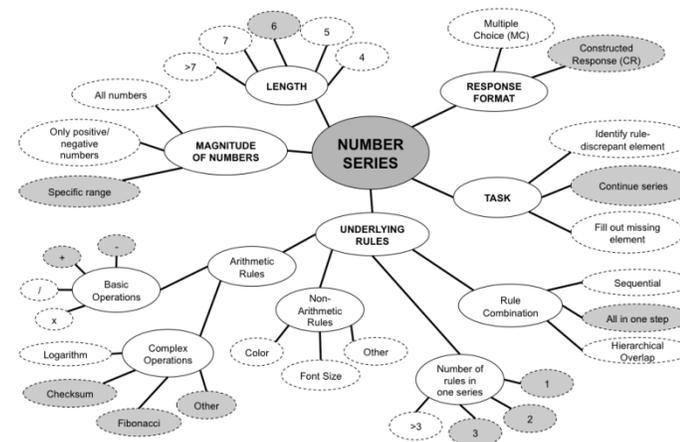
Multiple-Language AIG*

- Human translations add expense and error on top of an already expensive process of artfully crafted items
- Translated English-generated medical licensing examination multiple-choice items into Canadian French and Chinese by adding a “language layer” to the item models
- Still partly a work in progress since the “art” of translation is seldom exact, given the nuances of language
 - ◆ Sin embargo... el "arte" de la traducción es raramente exacto
 - ◆ However... the 'art' of the translation is rarely accurate

* Gierl, Lai & Turner (2012). Medical Education. Gierl, Fung, Lai & Zheng (2013). National Council on Measurement in Education Symposium Paper.

Rule-Based AIG for Number Series Items (J.P. Bertling, NCME, 2013)

- AIG model premise: *number series problems* are a convenient format to measure some aspects of numerical reasoning (application of rule-based induction) and are amenable to algorithmic item design
- There is a *mature* “task model” for number series problems



Standards for AIG

(Embretson & Poggio, NCME, 2013)

- AIG → less human involvement (\$\$\$)
- AIG without STRONG quality controls and evaluation criteria is not fruitful
- Standards that depend on projected use and quality of evidence
 - ◆ Quality of item content
 - ◆ Predictability of item parameters
 - ◆ Impact of item predictability on score reliability
- How much should traditional “content blueprints” drive these standards?

An Cognitive-Engineering Approach to Test Development

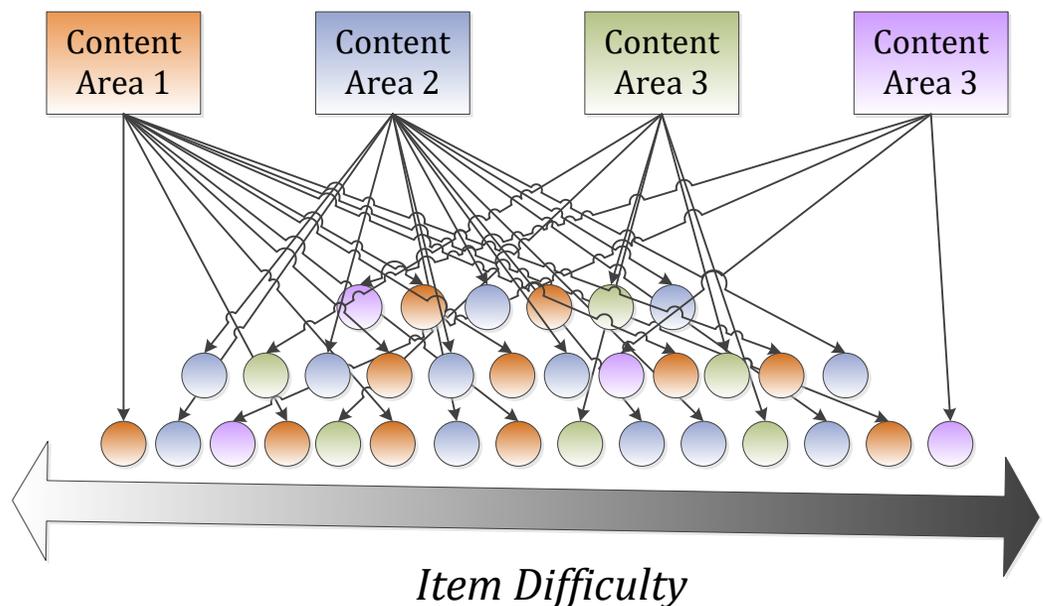
Assessment Engineering (AE)
combines the *scalability and replicability of AIG* (i.e., as an item-production mechanism) with *empirically verified cognitive task modeling and strong statistical quality controls*...all required for **isomorphism within ITEM FAMILIES**

**From a task modeling
perspective, content will NOT
necessarily be the same across
the SCALE because task models
differ in cOmplexity**

**Lower level skills
applied to simple
content**

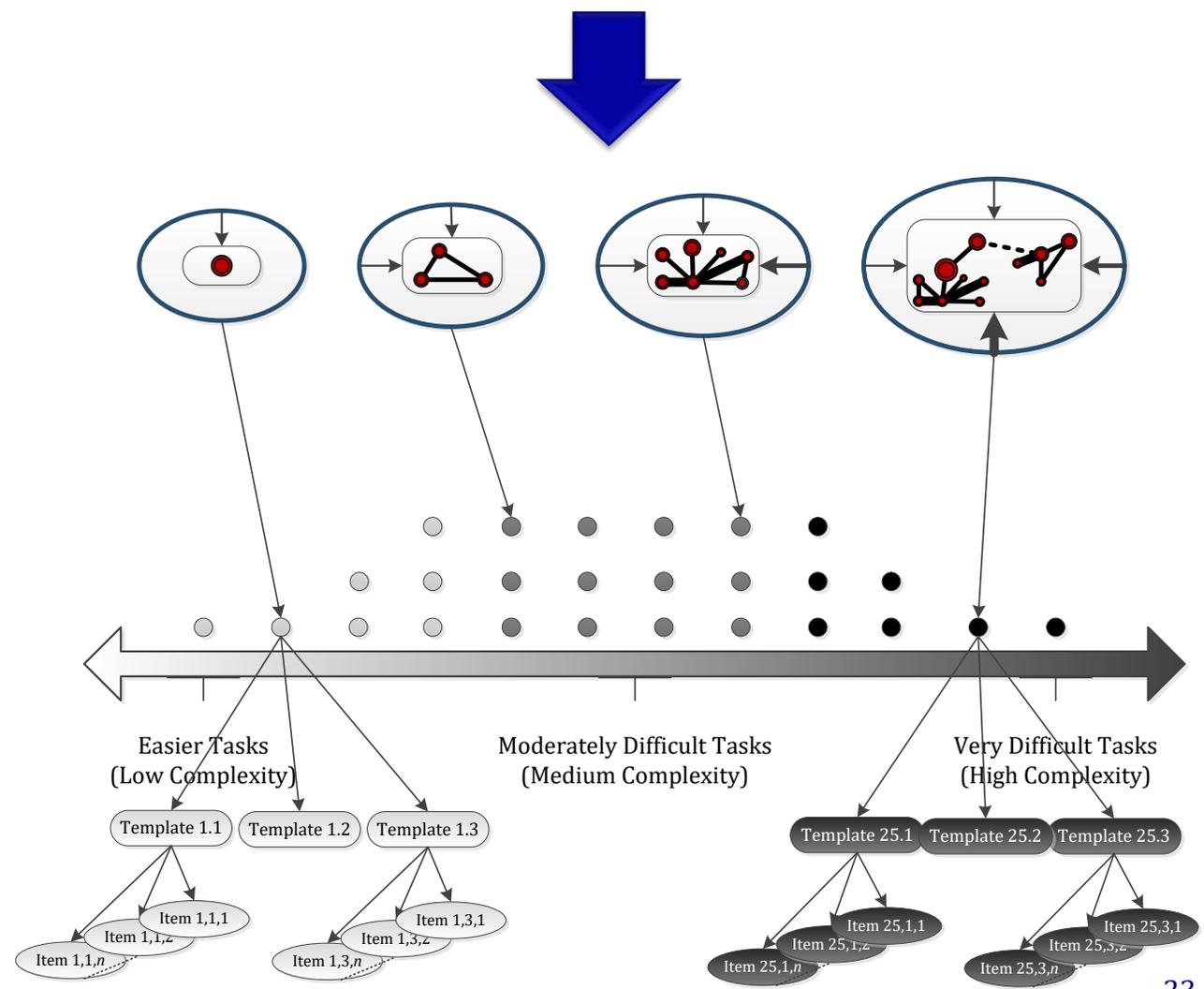
**High-level skills
applied to
complex content**





Traditional Item Writing and Test Assembly

Items as Part of a Task Model Family



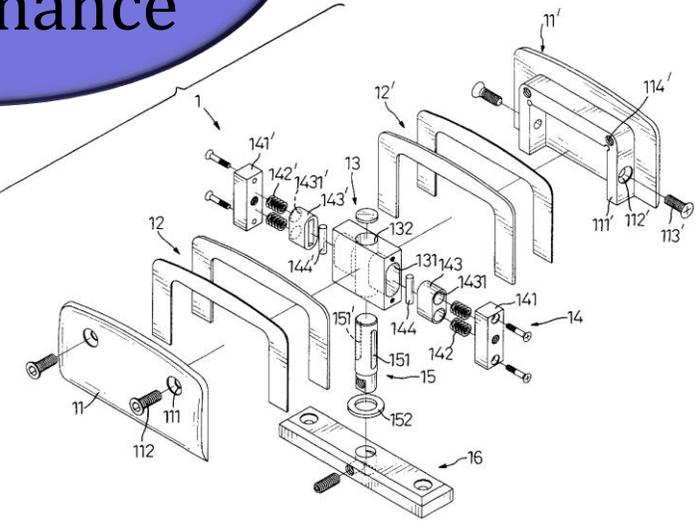
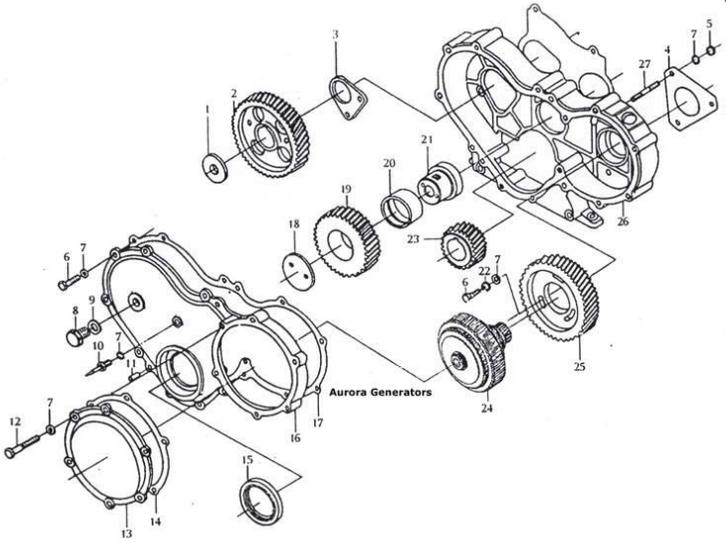
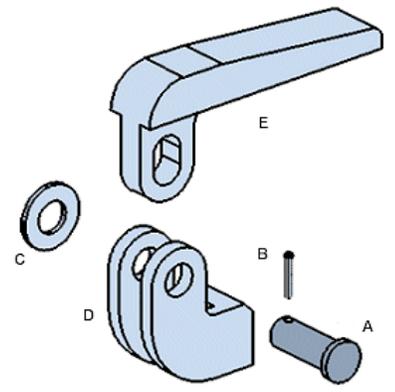
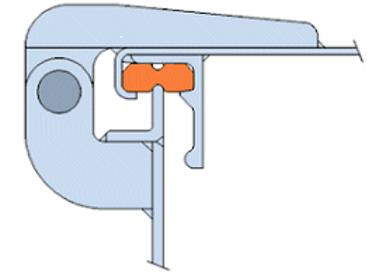
The Need for Engineering Principles Like Robust *Composability* to Item-Template Design

Standardized Components

Scalable & Replicable Designs

Stable Cross-Platform Performance

Consistent Data

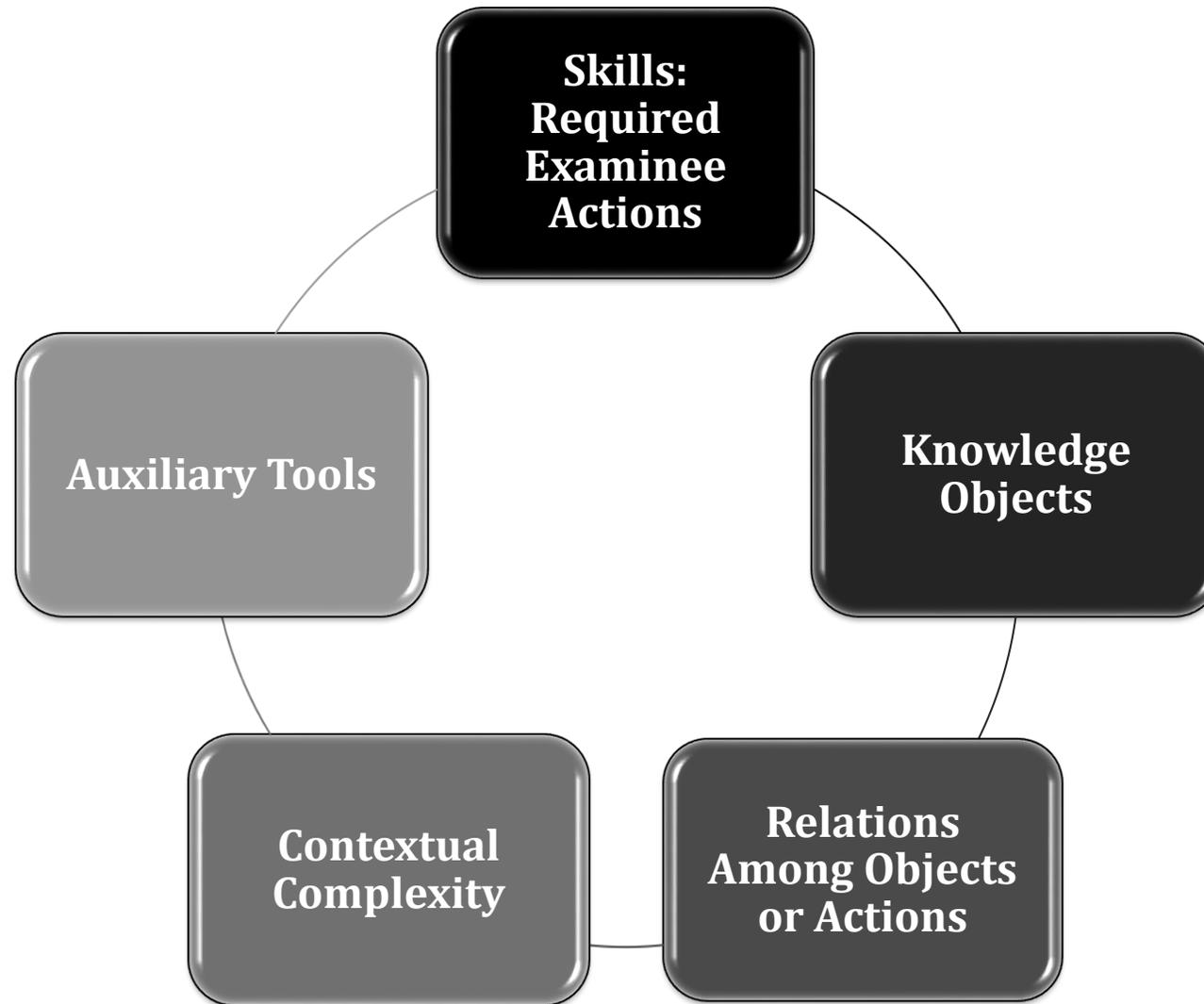


Cognitive Task Modeling

- **Task Model Grammars (TMGs)** are *domain-specific languages* that describe the intended **cognitive complexity design features** for families of assessment tasks—the **Task Models**
 - ◆ Content and declarative knowledge components
 - ◆ Procedural skills needed
 - ◆ Tools, resources
 - ◆ Contextual conditions
- **Task Model Maps (TMMs)** provide a distribution of Task Models on a scale

Cognitive Skill-Based Task Models

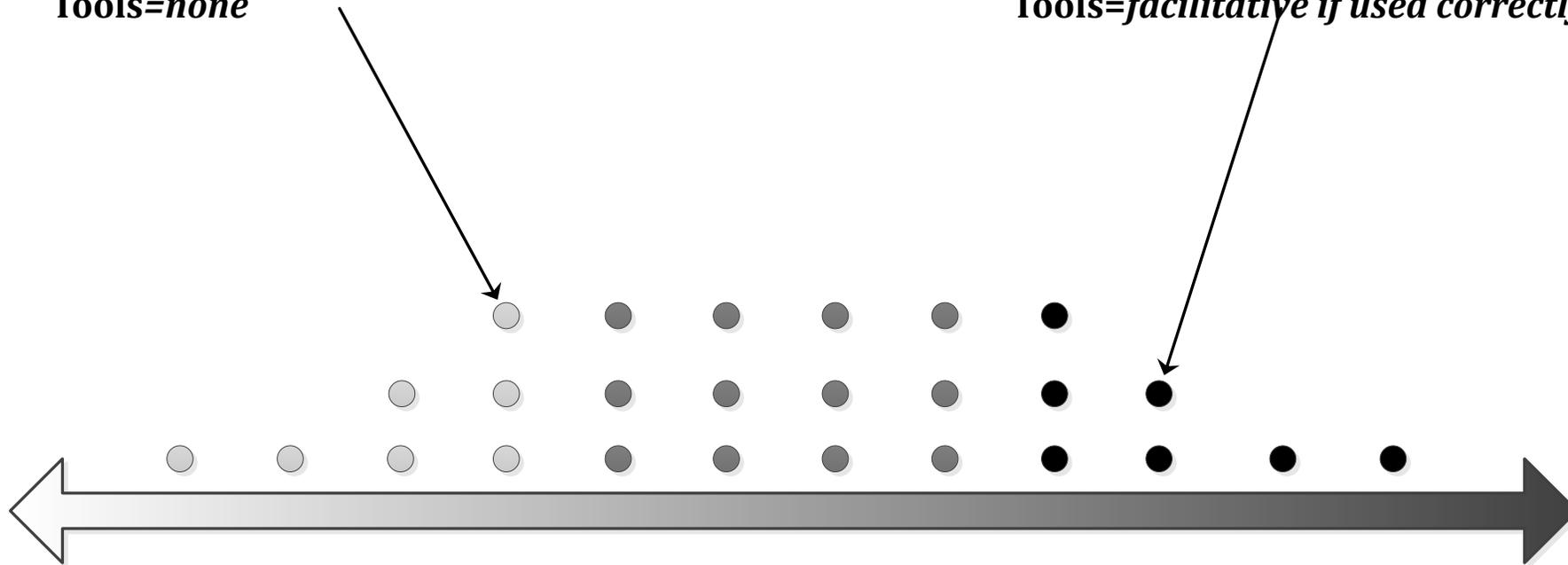
$action_2 [action_1 (is.related (object_1, object_2), object_3 | context, aux.tools)]$



Task Model Mapping: *Locating Intended Challenges* to Support Evidence-Based Claims

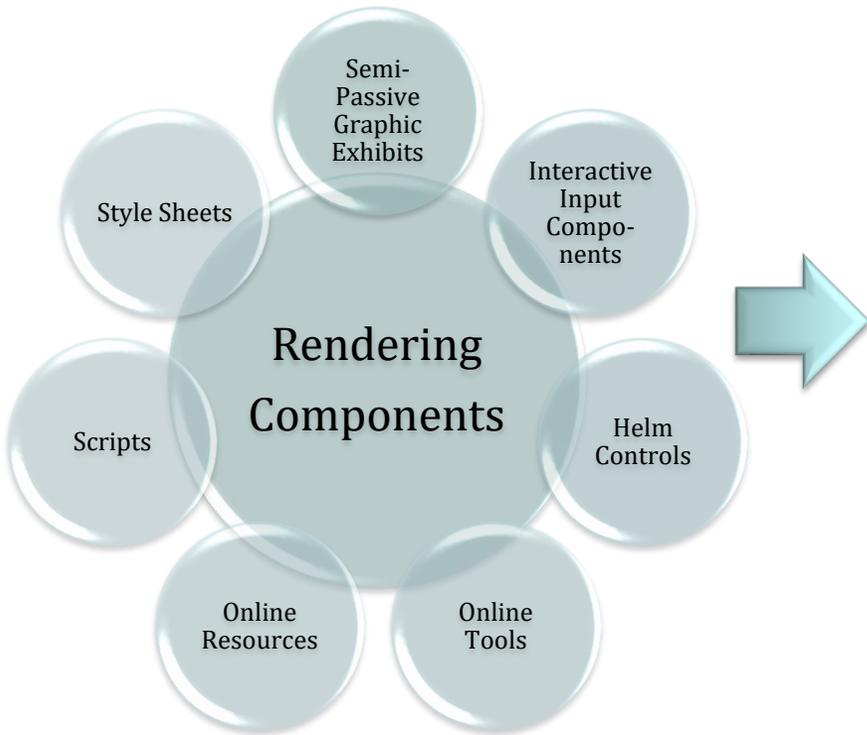
Skill=identify
Objects = one, simple concept
Relations=none
Context=match word → definition
Tools=none

Skills=identify, compare, evaluate
Objects = 3-4 complex properties
Relations=hierarchical (3 levels)
Context=complex text, dense info.
Tools=facilitative if used correctly

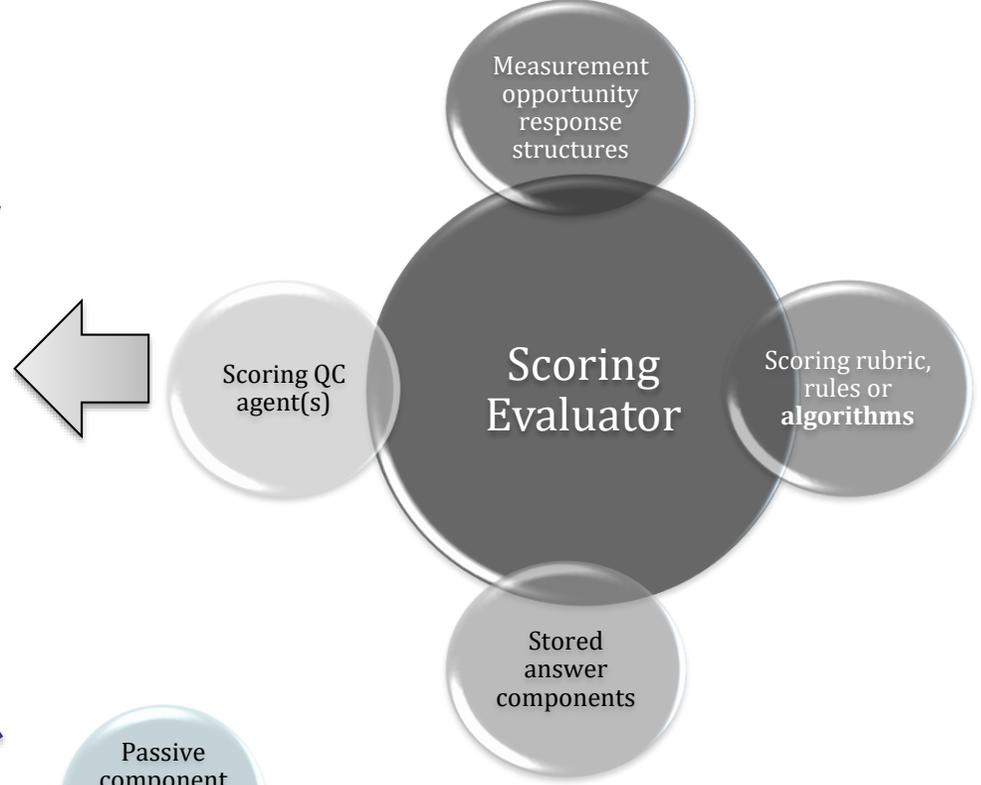


Types of Task Models

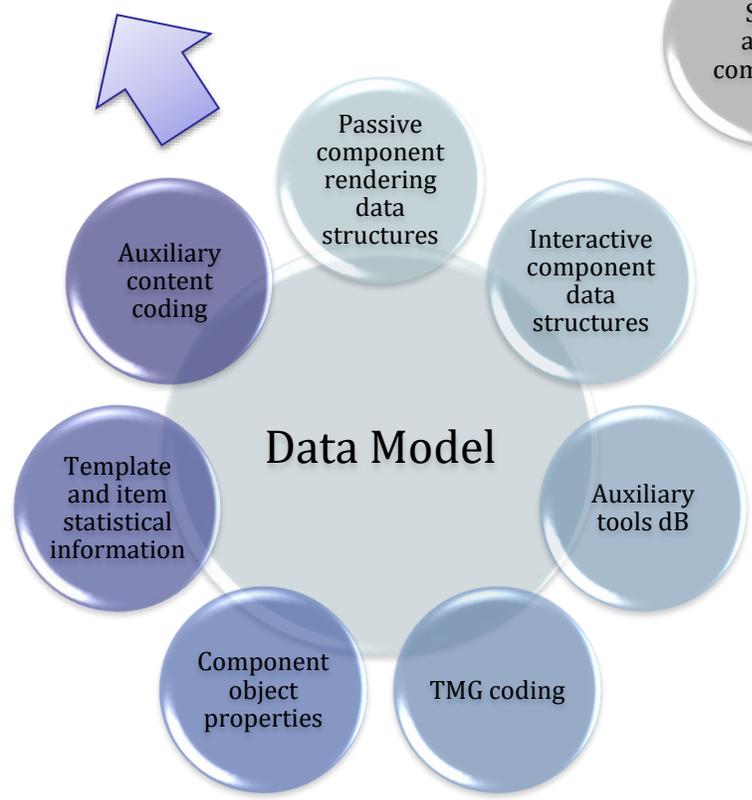
- ◆ ***Fixed-specification*** task models
 - ◆ Number of task models = no. of test items
 - ◆ Each task model is essential
- ◆ ***Domain-sampled*** task models
 - ◆ Multiple task models per location
 - ◆ Task models are considered exchangeable at a particular location
- ◆ ***Self-adaptive performance*** task models
 - ◆ Template components are manipulated to change the information (location)
 - ◆ Optimal reconfiguration of components



Hundreds of "item types" may be possible



Open-Ended Template Architecture for ANY Item Type



A Prototype Item for the CCSS H.S. Statistics & Probability Standard

Calculate expected values and use them to solve problems: (S-MD.3.) Develop a probability distribution for a random variable defined for a sample space in which **theoretical probabilities** can be calculated; find the expected value. (*CCSS Initiatives Project, www.corestandards.org/the-standards/mathematics/hs-statistics-and-probability/*)

A test has five multiple-choice questions scored correct/incorrect. Each question has four possible options. What will be the expected number-correct score for students who guess the answers to all five of the questions?

- A. 0.25
- B. 0.80
- C. 1.25
- D. 3.75
- E. 5.00

A Possible Set of TMG Specifications for Our Statistics $E(y)$ Standard

Calculate expected values and use them to solve problems: (S-MD.3.) Develop a probability distribution for a random variable defined for a sample space in which **theoretical probabilities** can be calculated; find the expected value. (*CCSS Initiatives Project*, www.corestandards.org/the-standards/mathematics/hs-statistics-and-probability/)

Recall.formula.SRS_uniform.discrete $\left[p_i = P_i(u_i = 1|a) = 1/a \right]$

Recall.formula.expected_value $\left[E(y) \doteq \bar{y} = \sum_{i=1}^n p_i u_i \right]$

Apply.formula.sum_products $\left[\bar{y} = \sum_{i=1}^n p_i u_i = p_1 u_1 + p_2 u_2 + \cdots + p_n u_n \right]$

Apply.formula.simplify_distributive $\left[\sum_{i=1}^n p u_i = p n \mid p = 1/a \right]$

Constraint.value.discrete_int $\left[u_i \in (0, 1, \dots, u^{\max}) \right]$

Constraint.value.discrete_int $\left[n \in (2, \dots, n^{\max}) \right]$

Constraint.value.prob $\left[0.0 \leq p \leq 1.0 \right]$

A Rendering Template for Our Simple Statistics Problem

A *<sample.event>* has *<n>* *<description.sample_units>*
<description.auxiliary_info>. *<The/Each>*
<description.theoretical_event_probability>. What will be the
expected *<description.value_unit(s)>* for
<description.objects_using_theoretical_prob_distrib>?

<MCq5.distractor.1=p>

*<MCq5.distractor.2=(1/n)*a>*

<MCq5.distractor.3= $n \cdot p = \sum_x x \cdot p_x$ >

<MCq5.distractor.4= $(1/a) \sum_x x = p \sum_x x$ >

*<MCq5.distractor.5=(1/a)*p*n>*

Scoring Evaluator

$u_i = \text{CAK}(i.\text{Selection.MCq.d} =$
 $i.\text{Key}, 1 \text{ if } T, 0 \text{ if } F)$

Note: $p = \text{theoretical_prob_distr.constant} = 1/a$

Approaches to Task Modeling

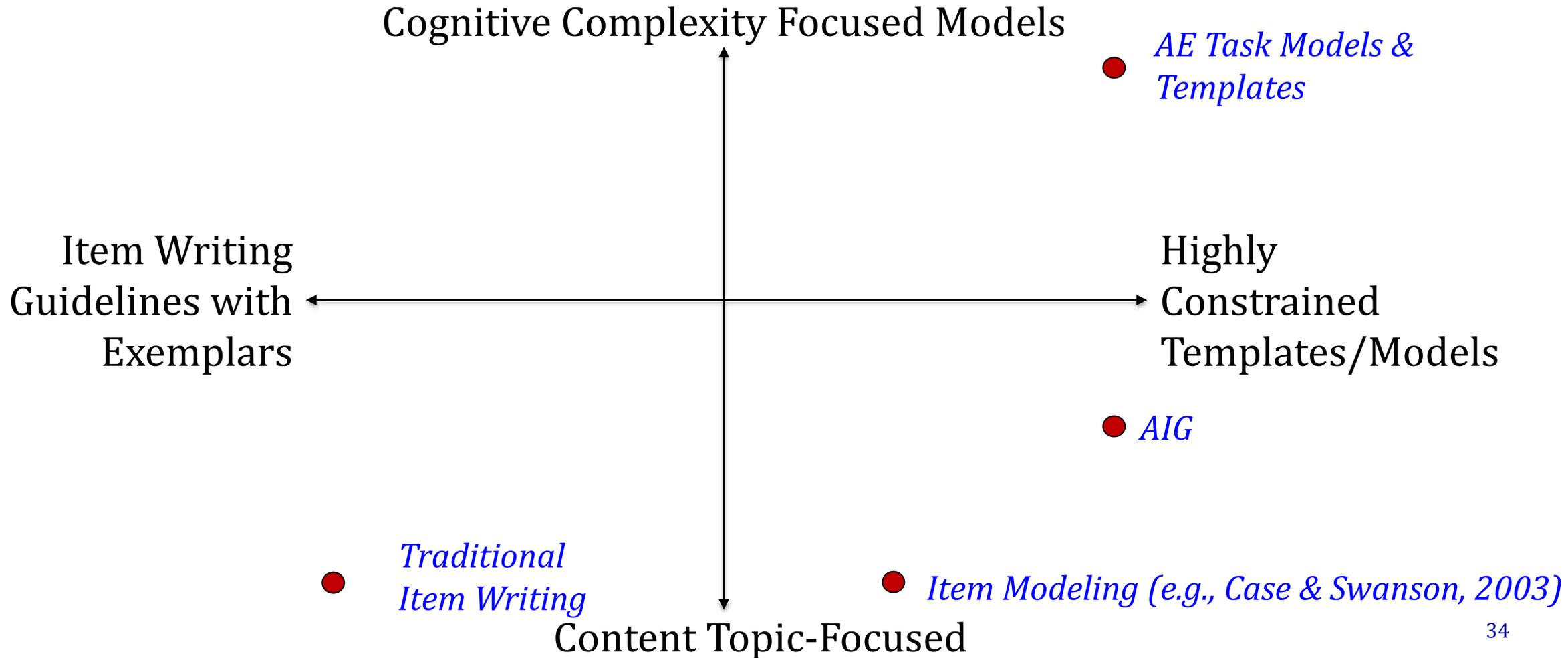
Reverse-Engineering Existing Items (Bottom-Up)

- Reverse engineering actual items to develop a TMG (propositional) or *language* to detail required skills and knowledge components
- Forward engineer templates and items from the TMG
- Iteratively refine of TMG-based families, matching empirical item difficulty ordering

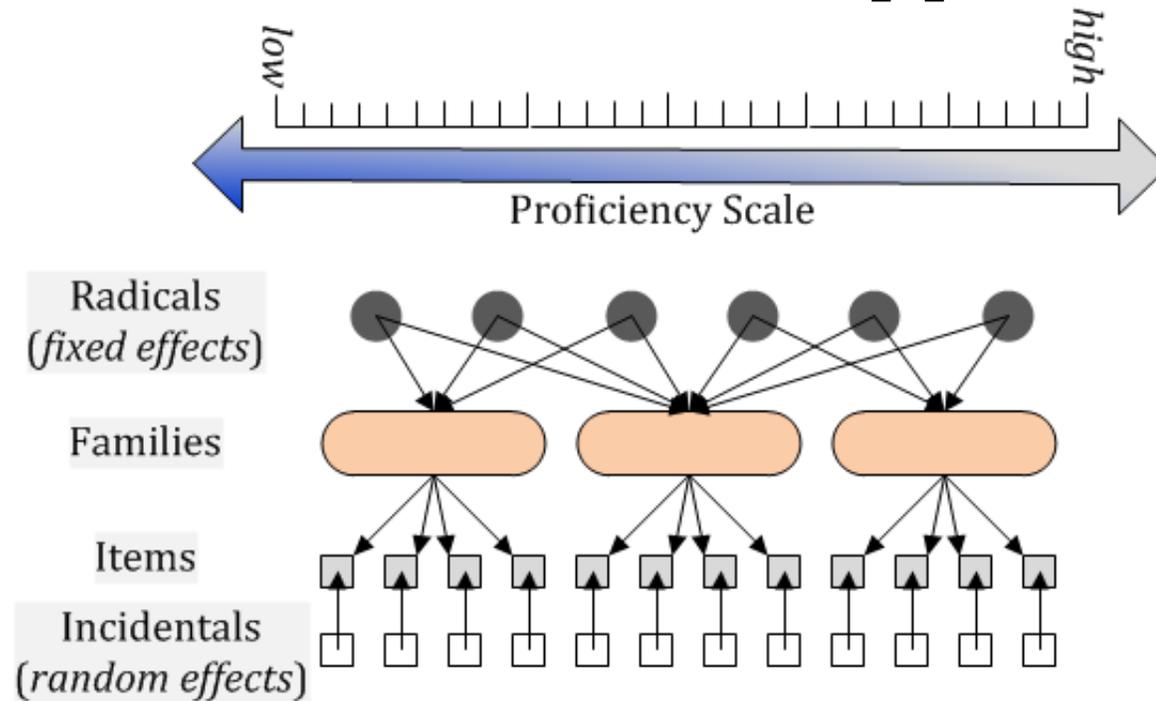
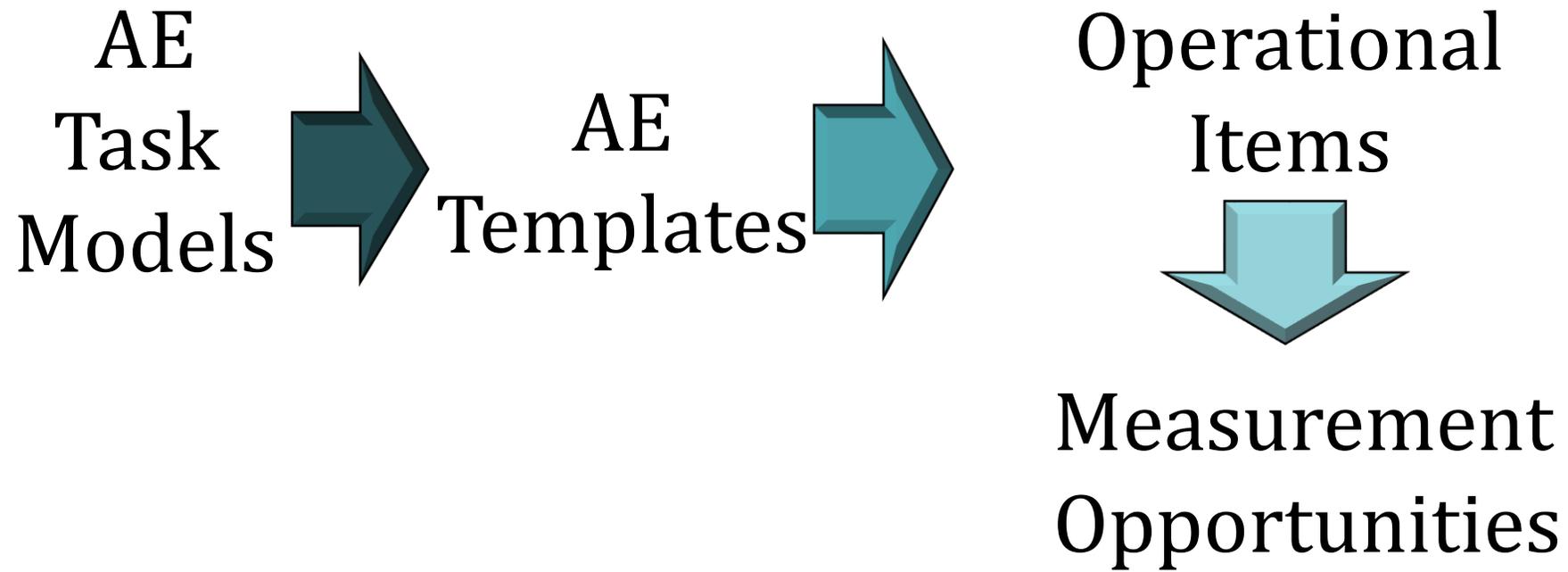
Construct Mapping Approach (Top-Down)

- Develop a TMM along a trajectory
- Design cognitive task models using challenge schema where *skills* \rightarrow *knowledge* | *context, tools*
- Iteratively design and validate templates and item families using a hierarchical QC approach

Item Models vs. Automatic Item Generation (AIG) vs. AE Task Models + Templates



Quality Control for Task Models and Templates: Items as True Item Families



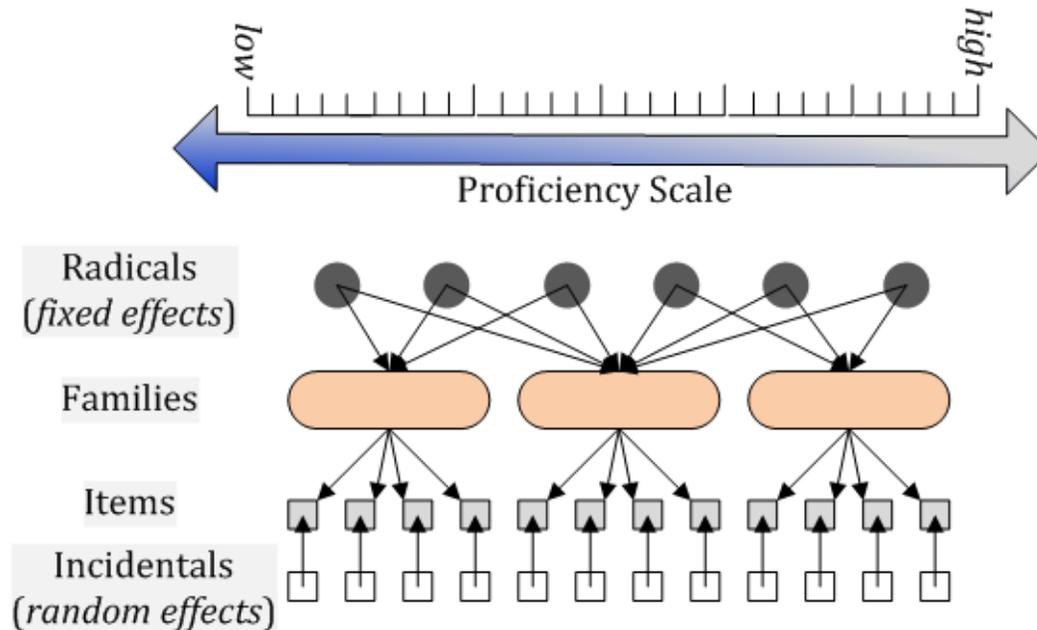
First-Level Model

$$P(x_{if} = 1 | \theta; \xi_{if}) = c_{if} + (1 - c_{if}) \Phi[a_{if}(\theta - b_{if})]$$

Second-Level Model

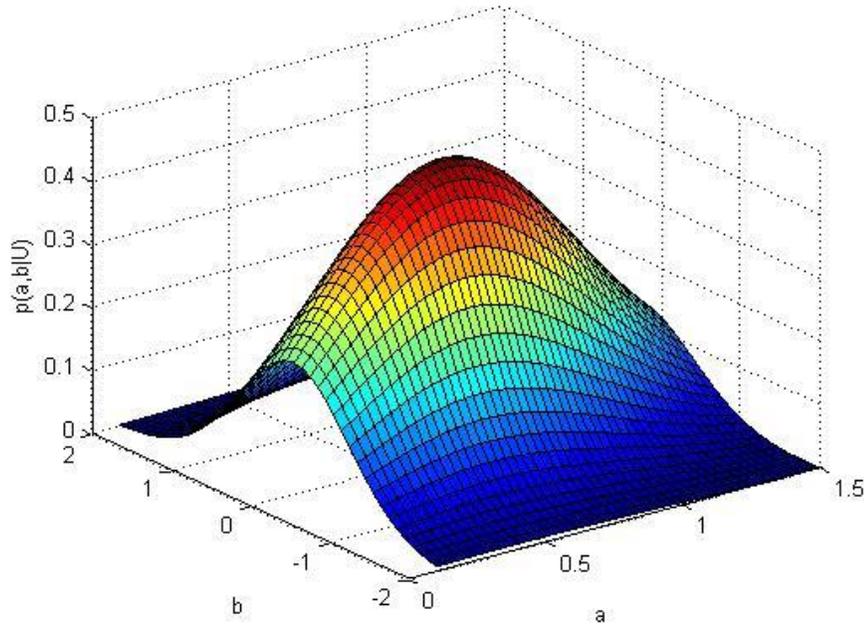
$$\xi_{if} = MVN(\mu_f, \Sigma_f)$$

$$\mu_{if} = \sum_{r=1}^R d_{fr} \beta_r$$

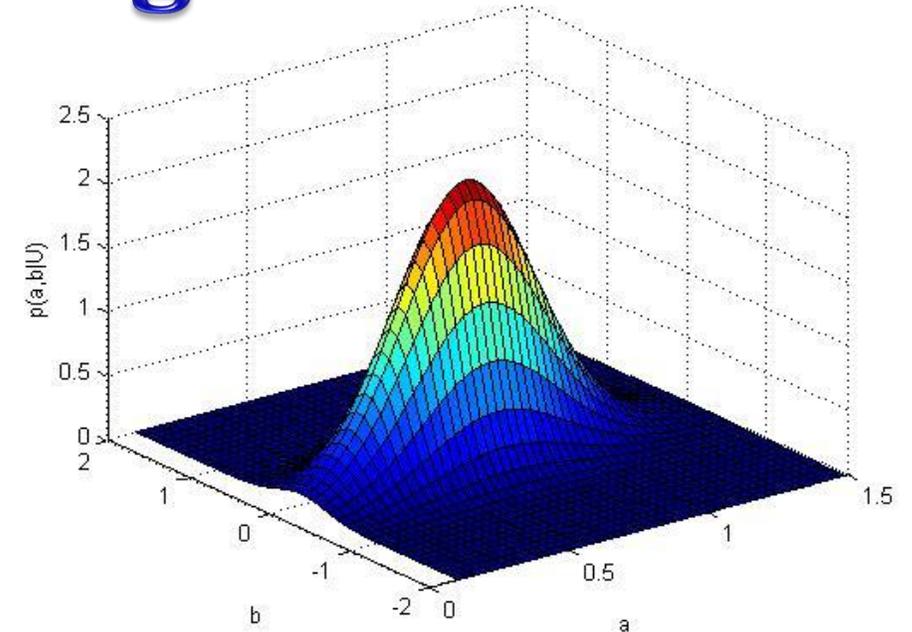


Geerlings, H., Glas, C. A. W., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, 76, 337-359.

QC and Variance of Item Parameter Estimates Drives Calibration Strategies

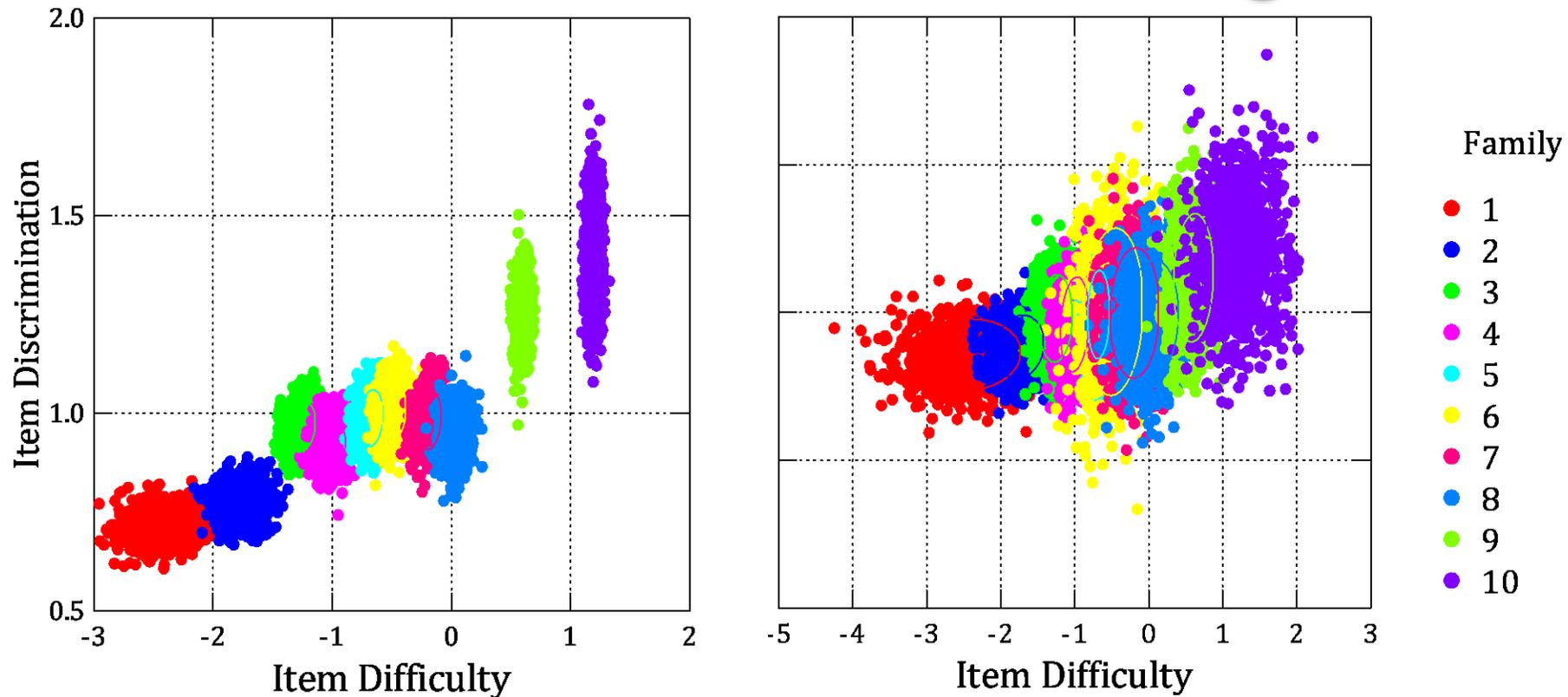


Calibrate individual items (ignore the item family)...Refine templates to reduce variation in item characteristics



Calibrate task models as families...monitor variation over time and “tweak” templates as needed

ESTIMATION and QC



QC Implications

- (a) Item families are working well
- (b) Calibrate TM families

QC Implications

- (a) Item families not working
- (b) Tighten/repair templates



**Psychometrics becomes part
of an continually active QC
system aimed maintaining
ROBUST SCALES using
hierarchically calibrating task
models or templates rather
than individual items**

Conclusion: Two Essential Conditions for the Success of AIG and AE

- ◆ **Substantive isomorphism** within item families
 - ◆ Cognitively exchangeable tasks in terms of required knowledge and skills
 - ◆ Exchangeable evidence to inform measurement claims
- ◆ **Statistical isomorphism** within item families
 - ◆ Sufficiently small variation of all item statistical properties within item families
 - ◆ Exchangeability of items within families for scoring purposes

Questions?

- ◆ “The world is full of magical things patiently waiting for our wits to grow sharper.”
 - ◆ Bertrand Russell

Thank you very much for your attention!

Ric Luecht: rmluecht@gmail.com

Matt Burke: mburke@abim.org