Reconceptualizing Items: \textit{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.”

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”
Different Item Generation Perspectives

Task Model Families

Population Distribution of Examinee Proficiency

Task Model Map for the Item Bank

N-Layered AIG & Mass Item Production Models

Easy Task Model Family

Moderate Task Model Family

Difficult Task Model Family

Item Shells

Item Writing/Editing as a Craft?

Task Model

Item Template

Item Model Variables

Simplify the following in equation:

\[ \log(10^{x} + 10^{y} - 10^{z}) \]

Elements

- 1 Value Range: 2 – 5 by 1
- 2 Value Range: 2 – 3 by 1
- 3 Value Range: 2 – 5 by 1
- 4 Value Range: 2 – 5 by 1

Solve (low complexity function)

several number of operations,

explicit application: three unique variables

<Patient.article><Patient.description.age>

"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> "?"
Where is the parietal lobe? (Click on the appropriate area of the brain, below.)

What is an “Item”?
Complications of Item Types

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

Item Types

- Multiple-Choice and Selected-Response
- Constructed- and Extended Response
- Simulations
- Technology-Enhanced
- Performance Assessments
- Short-Answer
- Multiple-Choice and Selected-Response
- Constructed- and Extended Response
- Simulations
- Technology-Enhanced
- Performance Assessments
- Short-Answer
Unconstrained

Actual Performance

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

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

True-False or Binary Choice Items

Constrained

Work Samples

Short-Answer (SA) Items

Multiple-Choice (MC) Items

Interactive Simulations

Technology-Enhanced (TE) Items

ARE MULTIPLE CHOICE EXAMS AN ACCURATE MEASURE OF ONE’S KNOWLEDGE?
A. YES
B. A AND C
C. A AND B
D. ALL OF THE ABOVE
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.

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

## AE Item Production

<table>
<thead>
<tr>
<th>Item Template #</th>
<th>Element 1</th>
<th>Element 2</th>
<th>Element 3</th>
<th>Element 4</th>
<th>Element 5</th>
<th>Number of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td></td>
<td></td>
<td>355</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td></td>
<td>762</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>9</td>
<td>33</td>
<td></td>
<td></td>
<td>5940</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>896</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>3</td>
<td>13</td>
<td>11</td>
<td></td>
<td>2132</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td></td>
<td>128</td>
</tr>
</tbody>
</table>

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

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 $\rightarrow$ 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.
Items as Part of a Task Model Family

Traditional Item Writing and Test Assembly
The Need for Engineering Principles Like Robust *Composability* to Item-Template Design

- Scalable & Replicable Designs
- Standardized Components
- 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

\[ action_2 \left[ \text{action}_1 \left( \text{is.related} \left( \text{object}_1, \text{object}_2 \right) \right), \text{object}_3 \mid \text{context}, \text{aux.tools} \right] \]
Task Model Mapping: *Locating Intended Challenges* to Support Evidence-Based Claims

**Skill**= identify  
**Objects** = one, simple concept  
**Relations**= none  
**Context**= match word $\rightarrow$ 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 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
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 \[ p_i = P_i \left( u_i = 1 \mid a \right) = \frac{1}{a} \]

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

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

Apply.formula.simplify_distributive \[ \sum_{i=1}^{n} pu_i = pn \mid p = 1 / a \]

Constraint.value.discrete_int \[ u_i \in (0,1,\ldots,u_{\text{max}}) \]

Constraint.value.discrete_int \[ n \in (2,\ldots,n_{\text{max}}) \]

Constraint.value.prob \[ 0.0 \leq p \leq 1.0 \]
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*p=\sum x*p_x>
<MCq5.distractor.4=(1/a)\sum x = p\sum x>
<MCq5.distractor.5=(1/a)*p*n>

Note: p=theoretical_prob_distr.constant=1/a

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

- Content Topic-Focused
- Item Writing Guidelines with Exemplars
- Traditional Item Writing
- Item Modeling (e.g., Case & Swanson, 2003)
- Highly Constrained Templates/Models
- AE Task Models & Templates
- Cognitive Complexity Focused Models

- AIG
Quality Control for Task Models and Templates: Items as True Item Families
AE Task Models $\rightarrow$ AE Templates $\rightarrow$ Operational Items

Measurement Opportunities

Proficiency Scale:
- Radicals (fixed effects)
- Families
- Items
- Incidentals (random effects)
First-Level Model

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

Second-Level Model

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

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

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
**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