Applicant reactions to AIG: A CAT
AIG feasibility study

Alan D. Mead\textsuperscript{1}, Sheng Zhang\textsuperscript{2}, Daniel Stopka\textsuperscript{2}

\textsuperscript{1}Talent Algorithms Inc.
\textsuperscript{2}Illinois Institute of Technology
Agenda

• Background & Introduction
  – Current Use and Applications of CAT AIG
  – Weak vs. Strong AIG
  – The Impact of Flawed Items

• Methods

• Results

• Discussions & Next Steps
How to build a CAT that uses AIG?

• Automatic item generation (AIG)
  – Our method seems promising (see Mead, 2013)

• Strong AIG
  – Generate items with known difficulty
  – We are working on this

• Avoid flawed items
  – AIG CAT items will be shown to examinees without prior review
  – AIG algorithm must avoid (or detect) flawed items
  – The current study
Weak vs. Strong AIG

• Weak AIG
  – Emphasis on generating different items
  – Items generated from a template
  – Little known about item difficulty

• Strong AIG
  – Emphasis on understanding cognitive process underlying responding
  – Strong theoretical (or empirical) model of item difficulty
  – Generate items from “scratch” based on strong theory/model (Embretson, 1999)

• AIG CAT requires strong AIG
Our AIG Method: Sample Item

Hat:Head
a) Blowgun:Dart
b) Mitten:Hand
c) Candy:Sweet
d) Neck:Necklace

• Identify a “bridge”; you wear HAT on HEAD
• Find a matching answer; you wear MITTEN on HAND
Our AIG Method: Generation

Hat:Head
a) Blowgun:Dart
b) Mitten:Hand
c) Candy:Sweet
d) Neck:Necklace

• Assemble a database of “bridges” with multiple pairs of words matching the bridge
• Sample two word pairs for stem and key
• Generate distractors
  • A work in progress
  • May be pairs from unrelated bridges
  • May be be pairs from same bridge manipulated to be incorrect
  • May be related words not matching the bridge
Current Study

• Completely avoiding flawed items will be hard
• This study seeks to understand examinee perceptions of flaws in items
  – Asked examinees to flag flawed items
  – A next step will be to evaluate the psychological effects of flawed items (on examinee exam perceptions and performance)
Study Goals

• Goal 1: Understand examinees’ perceptions of flawed items
  – Verify that “normal” AIG items are not perceived as flawed by examinees
  – Verify that items we have manipulated to be flawed are perceived as flawed by examinees

• Goal 2: Estimate psychometrics of AIG items
Method

• Participants:
  – N=33 recruited from MTurk
  – Final sample N=23 after data cleaning
Method

- 78 AIG items
  - 60 “normal” items generated from bridge item file
  - Types of distractor
  - 18 items manipulated to be flawed
- 21 “Applicant” reaction items
We considered six possible flaws:

1. Two Correct Keys
2. No Correct Keys
3. One Gibberish Distractor
4. Extremely Difficult Word Sense
5. Trivially Easy (Due to Flawed Generation)
6. Stilted Analogy

Wrote three items for each flaw.
RESULTS
AIG Difficulty

• For all the items:
  – Mean = .668; SD = .179

• For items that were not being flagged:
  – Mean = .694; SD = .183
Which items were flagged?

• 18 items were manipulated to have flaws
  – 3 items for each of the six types of flaws
  – 14 (78%) were flagged by 1 or more examinees
    • Median proportion = 16% (i.e., 4 people)
  – 4 (12%) were not flagged (failed as flawed items)

• 60 AIG items (hopefully not flagged)
  – 36 (60%) were flagged by 1 or more examinees
    • Median proportion = 4% (i.e., one person)
  – 24 (40%) were not flagged (succeeded as “normal” items)
Flagging by type of flaw

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
<th>Not Flagged</th>
<th>Prop. Flagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two correct keys</td>
<td>3</td>
<td>2</td>
<td>0.01</td>
</tr>
<tr>
<td>No correct keys</td>
<td>3</td>
<td>0</td>
<td>0.33</td>
</tr>
<tr>
<td>One gibberish distractor</td>
<td>3</td>
<td>0</td>
<td>0.15</td>
</tr>
<tr>
<td>Difficult word sense</td>
<td>3</td>
<td>1</td>
<td>0.09</td>
</tr>
<tr>
<td>Trivially easy</td>
<td>3</td>
<td>1</td>
<td>0.06</td>
</tr>
<tr>
<td>Stilted analogy</td>
<td>3</td>
<td>0</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Discussions

• Overall, examinees flagged more proportions of intended flawed items than unflawed items.
• Some types of flawed items were more likely to be flagged than the others.
• 10 examinees (30%) were excluded from the study due to not paying enough attention or not putting enough effort into answering
Next Steps

• Sufficient flawed items do seem to be perceived by participants.
• Conditions with different percentages of flawed items will be tested and compared.
• Order of flawed items in the test. Flawed items will appear early in the test.
• Participants’ test-taking motivation.
Thank you!

amead@alanmead.org
szhang103@hawk.iit.edu
dstopka@hawk.iit.edu