Extending the Role of Psychometrics in Integrated Security Plans

Lisa S. O’Leary, Ph.D.

Alpine Testing Solutions
Context

• Rampant test fraud in information technology (IT) certification testing programs
• Widespread unauthorized exposure of exam forms and perpetual item exposure
• Continually administered exams
• Tight timeframes for piracy (days or weeks)

...leads to a need for realistic approaches to exam maintenance that can identify specific compromised content to reduce threats to the validity of score interpretation and use
Goals

• Address rampant test fraud in IT certification
• Minimize the costs of piracy
• Maximize the measurement integrity and validity
• Deter future test fraud

Well-articulated data forensic procedures are “a very effective way to communicate to candidates that cheaters leave behind irregular patterns of responses, and that even if they are sufficiently clever to successfully cheat on the exam, they will be unearthed by sophisticated statistical procedures being run in the background” (Wollack & Fremer, 2013, p. 11)
Goals

- Proactive consideration of security throughout the test development process
  - Collecting continual security evidence supports the intended use/interpretation of test scores and integrity of credentialing decisions
  - Continually addressing program design, legal considerations, content development, and psychometric analyses protects the security of exam content
Credential Program 1

• Test available to everyone
• On-demand administration
• Immediate score report
  – Automatic award
• Six month health check
• Annual Update

Reactionary Posture

• Feedback
  • Stakeholders like convenience
  • Greater respect for earlier awards
  • Limited view of data
  • Delayed response to trends

• Award viewed with skepticism
• Confused why value is decreasing
Credential Program 2

- Test available to everyone
- On-demand administration
- Immediate score report
  - Automatic award
- Monthly security report
- Six month health check
- Periodic updates

Feedback
- Stakeholders like convenience
- Greater respect for earlier awards
- Award viewed with skepticism
- Confused why value is decreasing
Credential Program 3

- Test available to eligible candidates
- Windowed administration
- Delayed score report (6 weeks)
- Security
- Pilot items
- Item selection and forms assembly each window
- Periodic major updates

Proactive Posture
- Identify problems prior to award
- Grow item bank
- Limited view of data
- Timely response to some problems

Feedback
- Stakeholders complain about limited windows and lag in reporting
- No credit for maintaining value

© 2013 Alpine Testing Solutions, Inc.
Credential Program 4

- Test available to eligible candidates
- On-demand administration
- Near real-time score report (48-72 hours)
  - Automatic award
- Pilot items in every administration
- Data repository
  - Daily security monitoring
  - Daily health monitoring
- Periodic updates

Proactive Posture
- Identify and address:
  - Growth
  - Trends
  - Feedback

Feedback
- Stakeholders like convenience
- No credit for maintaining value

© 2013 Alpine Testing Solutions, Inc.
Security Baker’s $\frac{1}{2}$ Dozen

1) Candidate eligibility
2) Protecting intellectual property
3) Candidate flagging criteria
4) Differential performance flagging criteria
5) Alignment of exam maintenance plan
6) Alignment of forms maintenance and retake policy
7) Candidate education
Candidate Eligibility

Initial requirement

- **Required: All**
  - **Required: 1**
    - [800-411] IT Education Core Competencies
  - **Required: 1**
    - IT Education Program Agreement
  - **Required: 25**
    - IT Education Program Base Maintenance Fee - $25
    - IT Education Program CE Maintenance Surcharge - $100
    - IT Education Program Fee - CE Upgrade - $75

Eligibility Status

<table>
<thead>
<tr>
<th>Exam Name</th>
<th>Status</th>
<th>Eligibility ID Number</th>
<th>Earliest test date</th>
<th>Testing ends on</th>
<th>Register to test</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT Education Core Competencies</td>
<td>Active</td>
<td>002819610</td>
<td>02-25-2011</td>
<td>03-28-2016</td>
<td>Schedule exam</td>
</tr>
<tr>
<td>Robust TCP/IP Network Systems Architecture</td>
<td>Active</td>
<td>002819614</td>
<td>02-25-2011</td>
<td>03-29-2016</td>
<td>Schedule exam</td>
</tr>
<tr>
<td>Scholastic Learning Systems Architecture</td>
<td>Active</td>
<td>002742379</td>
<td>01-19-2011</td>
<td>03-28-2015</td>
<td>Schedule exam</td>
</tr>
<tr>
<td>Instructional Development</td>
<td>Used</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marketing 101</td>
<td>Used</td>
<td>038395475</td>
<td>03-24-2011</td>
<td>06-31-2079</td>
<td>Schedule exam</td>
</tr>
<tr>
<td>Functional Languages</td>
<td>Active</td>
<td>002819513</td>
<td>02-25-2011</td>
<td>03-28-2016</td>
<td>Schedule exam</td>
</tr>
<tr>
<td>Object Oriented Programming</td>
<td>Used</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Protecting IP

- Candidate agreements
- SME/Committee member agreements
- Document security
- Data security
Protecting IP

• Centralized Scoring
  • Keys stored in one, central location
  • Security checking prior to awarding grade/credential
  • Customized score reports
Candidate Flagging Criteria

- Set default flags to identify candidates with potentially suspect exam behavior or performance

<table>
<thead>
<tr>
<th>Security Flag</th>
<th>Suspect Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam Retakes</td>
<td>Candidate takes same exam ( x ) times within ( y ) period</td>
</tr>
<tr>
<td>Rapid Exam Completion</td>
<td>( x% ) percent of items completed in less than ( y ) seconds</td>
</tr>
<tr>
<td>Retake After Pass</td>
<td>Candidate takes the same exam after already passing</td>
</tr>
<tr>
<td>Large Score Differential</td>
<td>Score increase by more than ( x% ) within 2 attempts within ( y ) days</td>
</tr>
<tr>
<td>High score/Low time</td>
<td>Exam score above ( x% ), time spent on exam less than ( y ) min</td>
</tr>
<tr>
<td>Too Little Exam Time</td>
<td>Candidate spent less than ( x ) minutes taking an exam</td>
</tr>
<tr>
<td>Possible Collusion</td>
<td>Candidates at the same test center on same date and scored within ( y% ) of each other on same exam</td>
</tr>
<tr>
<td>Security Items</td>
<td>Candidate correctly answered ( x ) security items out of ( y ) total security items</td>
</tr>
<tr>
<td>Differential Item Performance</td>
<td>( x% ) or above on 1(^{st}) item type and ( y% ) or under on 2(^{nd}) item type</td>
</tr>
<tr>
<td>Watch List</td>
<td>Candidate is on the watch list at the time of taking a test</td>
</tr>
<tr>
<td>Banned list</td>
<td>Candidate is on the ban list at the time of taking a test</td>
</tr>
</tbody>
</table>

© 2013 Alpine Testing Solutions, Inc.
Candidate Flagging Criteria

- Leads to evidence of possible form exposure and/or candidate pre-knowledge during operational administration

© 2013 Alpine Testing Solutions, Inc.
Differential Performance

- Differential person functioning (DPF)
  - Identify candidates likely to have had prior knowledge of exam content
  - Expect candidates with prior content knowledge to have high ability on scored items and low ability on unscored items; low probability of the two measures resulting from the same candidate
Differential Performance

- Use aggregate results to identify trends across test centers or geographic locations with potential issues or suspect patterns
Differential Performance

• Differential item functioning (DIF)
  – Assess extent to which candidates’ item pre-knowledge impacts item performance
  – Determine degree of item degradation and gather information to drive exam maintenance
  – Compare performance of DPF-flagged candidates to DPF non-flagged candidates
  – Expect compromised items to favor candidates with item pre-knowledge (DPF-flagged candidates); non-exposed items to be of equal difficulty to both candidate subgroups
Differential Performance

• Practical considerations
  – Item bank size
  – Bank exposure rate
  – Ratio of scored to unscored items
  – Differential performance of scored and unscored items
  – Impact of item degradation
  – Availability of new content
  – Capacity for follow-up action
Differential Performance

- DPF in conjunction with DIF can be used to:
  1. detect when security breaches have occurred;
  2. determine the overall extent of item exposure;
  3. build cases against suspect candidates;
  4. collaborate with other evidence to support the enforcement of sanctions against candidates;
  5. highlight specific items with compromised content;
  6. evaluate appropriate next steps for particular items and entire item banks

...all while discussing the relevant psychometric and policy issues for each of these areas
Case Study

- 8,350 administrations of large-scale IT cert. exam
- 641 total items (227 scored, 414 unscored)
- Substantial item exposure issues
Case Study

• Compared candidates’ performance on scored (80) and unscored (20) items
  – Assumed only scored items were exposed; unscored items were not yet compromised

• 531 candidates (6.4%) flagged for DPF
Case Study

- 138 items (20.2%) displayed DIF

### Differential Item Functioning

<table>
<thead>
<tr>
<th>Item Status</th>
<th>Significant DIF</th>
<th>No DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scored</td>
<td>57</td>
<td>169</td>
</tr>
<tr>
<td>Unscored</td>
<td>0</td>
<td>334</td>
</tr>
</tbody>
</table>

© 2013 Alpine Testing Solutions, Inc.
Case Study

- 15 of the 169 scored items without evidence of DIF selected as anchor items for upgrade exam
  - Anchor items were proportionately representative of the blueprint; well-fitting to the model

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Section</th>
<th>Rasch Measure</th>
<th>P-value</th>
<th>Item-Score Correlation</th>
<th>DIF Measure</th>
<th>DIF S.E.</th>
<th>DIF Measure</th>
<th>DIF S.E.</th>
<th>DIF Contrast</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>0.61</td>
<td>0.60</td>
<td>0.53</td>
<td>-0.21</td>
<td>0.28</td>
<td>0.64</td>
<td>0.04</td>
<td>-0.85</td>
<td>0.003</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0.42</td>
<td>0.64</td>
<td>0.49</td>
<td>-0.14</td>
<td>0.27</td>
<td>0.42</td>
<td>0.04</td>
<td>-0.56</td>
<td>0.045</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>-0.33</td>
<td>0.77</td>
<td>0.33</td>
<td>-1.06</td>
<td>0.39</td>
<td>-0.33</td>
<td>0.05</td>
<td>-0.73</td>
<td>0.063</td>
</tr>
<tr>
<td>47</td>
<td>1</td>
<td>-0.27</td>
<td>0.76</td>
<td>0.33</td>
<td>-0.79</td>
<td>0.29</td>
<td>-0.27</td>
<td>0.05</td>
<td>-0.52</td>
<td>0.085</td>
</tr>
<tr>
<td>57</td>
<td>1</td>
<td>-1.43</td>
<td>0.90</td>
<td>0.27</td>
<td>-2.37</td>
<td>0.7</td>
<td>-1.43</td>
<td>0.07</td>
<td>-0.94</td>
<td>0.185</td>
</tr>
<tr>
<td>66</td>
<td>2</td>
<td>0.69</td>
<td>0.58</td>
<td>0.34</td>
<td>0.59</td>
<td>0.16</td>
<td>0.69</td>
<td>0.04</td>
<td>-0.1</td>
<td>0.560</td>
</tr>
<tr>
<td>79</td>
<td>2</td>
<td>-1.57</td>
<td>0.91</td>
<td>0.28</td>
<td>-1.95</td>
<td>0.59</td>
<td>-1.57</td>
<td>0.07</td>
<td>-0.38</td>
<td>0.520</td>
</tr>
<tr>
<td>102</td>
<td>2</td>
<td>-1.43</td>
<td>0.90</td>
<td>0.25</td>
<td>-3.04</td>
<td>0.89</td>
<td>-1.43</td>
<td>0.07</td>
<td>-1.62</td>
<td>0.073</td>
</tr>
<tr>
<td>130</td>
<td>3</td>
<td>-1.29</td>
<td>0.89</td>
<td>0.25</td>
<td>-1.06</td>
<td>0.39</td>
<td>-1.29</td>
<td>0.06</td>
<td>0.23</td>
<td>0.565</td>
</tr>
<tr>
<td>151</td>
<td>3</td>
<td>-0.18</td>
<td>0.74</td>
<td>0.53</td>
<td>-1.82</td>
<td>0.42</td>
<td>-0.14</td>
<td>0.05</td>
<td>-1.68</td>
<td>0.000</td>
</tr>
<tr>
<td>160</td>
<td>4</td>
<td>-0.16</td>
<td>0.74</td>
<td>0.44</td>
<td>-0.7</td>
<td>0.28</td>
<td>-0.16</td>
<td>0.05</td>
<td>-0.54</td>
<td>0.063</td>
</tr>
<tr>
<td>175</td>
<td>4</td>
<td>-0.4</td>
<td>0.78</td>
<td>0.52</td>
<td>-1.98</td>
<td>0.42</td>
<td>-0.36</td>
<td>0.05</td>
<td>-1.62</td>
<td>0.000</td>
</tr>
<tr>
<td>182</td>
<td>4</td>
<td>-0.35</td>
<td>0.77</td>
<td>0.39</td>
<td>-0.55</td>
<td>0.27</td>
<td>-0.35</td>
<td>0.05</td>
<td>-0.2</td>
<td>0.473</td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>-1.2</td>
<td>0.88</td>
<td>0.31</td>
<td>-2.36</td>
<td>0.69</td>
<td>-1.2</td>
<td>0.06</td>
<td>-1.16</td>
<td>0.093</td>
</tr>
<tr>
<td>212</td>
<td>6</td>
<td>-1.51</td>
<td>0.91</td>
<td>0.31</td>
<td>-2.07</td>
<td>0.51</td>
<td>-1.51</td>
<td>0.07</td>
<td>-0.55</td>
<td>0.283</td>
</tr>
</tbody>
</table>
Case Study

- Anchor items’ item difficulty estimates stable within upgrade item bank

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Section</th>
<th>Number of Responses</th>
<th>P-value</th>
<th>Rasch Measure</th>
<th>Displacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>404</td>
<td>0.59</td>
<td>0.61</td>
<td>0.08</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>411</td>
<td>0.62</td>
<td>0.42</td>
<td>0.13</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>388</td>
<td>0.73</td>
<td>-0.33</td>
<td>0.29</td>
</tr>
<tr>
<td>47</td>
<td>1</td>
<td>399</td>
<td>0.75</td>
<td>-0.27</td>
<td>0.14</td>
</tr>
<tr>
<td>57</td>
<td>1</td>
<td>424</td>
<td>0.92</td>
<td>-1.43</td>
<td>-0.3</td>
</tr>
<tr>
<td>66</td>
<td>2</td>
<td>414</td>
<td>0.58</td>
<td>0.69</td>
<td>0.02</td>
</tr>
<tr>
<td>79</td>
<td>2</td>
<td>397</td>
<td>0.91</td>
<td>-1.57</td>
<td>0.09</td>
</tr>
<tr>
<td>102</td>
<td>2</td>
<td>444</td>
<td>0.91</td>
<td>-1.43</td>
<td>-0.01</td>
</tr>
<tr>
<td>130</td>
<td>3</td>
<td>444</td>
<td>0.90</td>
<td>-1.29</td>
<td>-0.02</td>
</tr>
<tr>
<td>151</td>
<td>3</td>
<td>436</td>
<td>0.79</td>
<td>-0.18</td>
<td>-0.24</td>
</tr>
<tr>
<td>160</td>
<td>4</td>
<td>416</td>
<td>0.78</td>
<td>-0.16</td>
<td>-0.14</td>
</tr>
<tr>
<td>175</td>
<td>4</td>
<td>412</td>
<td>0.77</td>
<td>-0.4</td>
<td>0.13</td>
</tr>
<tr>
<td>182</td>
<td>4</td>
<td>401</td>
<td>0.79</td>
<td>-0.35</td>
<td>-0.06</td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>418</td>
<td>0.87</td>
<td>-1.2</td>
<td>0.19</td>
</tr>
<tr>
<td>212</td>
<td>6</td>
<td>395</td>
<td>0.89</td>
<td>-1.51</td>
<td>0.24</td>
</tr>
</tbody>
</table>
Exam Maintenance Plan

Conduct Health Check

- Administer Operational Forms that were built from Beta Item Selection Analysis
- Review exam and form-level statistics
- Delete items that are not performing well
- Set aside items viable after revision with SMEs
- Include well performing items on newly proposed forms
- Seed unscored items to pilot and obtain statistics

Forms Re-Assembly

- Administer newly proposed forms, equated to prior version
- Conduct health check
- Update item banking with revised statistics and item decisions (as needed)
- Seed unscored items to pilot and obtain statistics

Forms Re-Assembly OR Content Refresh

© 2013 Alpine Testing Solutions, Inc.
Exam Maintenance Plan

• Piloting unscored items on operational forms allows for:
  – Systematic content refreshing to enable retirement of items with poor statistical performance, suspected exposure issues, or expected bias
  – Proactive approach to enable addition of new items representative of content changes, upgrades, or updates

Pros

- Refresh exam content & expand available item pool for future forms
- Obtain item statistics on new items within an operational exam
- Candidate effort is high as all items appear to be scored
- Compare candidate performance on scored versus unscored items

Cons

- Need a sufficiently large item pool to have unscored items to pilot
- Risk exposure of unscored pilot items
- Can add additional time burden to existing forms
Exam Maintenance Plan

• Rapidity of content shifts impacts the frequency of necessary analysis and maintenance
  – Exams in dynamic and quickly changing domains require more frequent maintenance than those in more static domain areas

• Likelihood or suspicion of suspect candidates or exam behavior impacts the frequency of necessary analysis and maintenance
  – Compromised exams require more immediate maintenance to gauge the impact of the security breach
Forms Maintenance & Retakes

- Form A: Initial Attempt
- Form B: Retake Attempt
- Form C: Retake Attempt
- Form D

Forms Refresh after 6 Months
Candidate Education

- Candidates are your friends!
- Value proposition of program
- Self policing
Conclusions

• Proactive consideration of security throughout the test development process increases validity of candidate decisions and the testing program.

• Data repository enables diversified approach to exam security:
  – Timely candidate detection and enforcement
  – Routine tracking and exam maintenance over time
  – In-depth analyses to address specific concerns

• Probabilistic-based methods for detection of suspect candidates enhance defensibility of actions and enforcement.
Contact Information

• Lisa S. O'Leary, Ph.D.
  Psychometrician
  lisa.oleary@alpintesting.com

www.alpintesting.com