Overview of Presentation

1. Legal Foundation of Data Forensics
   - Jennifer Ancona Semko
2. Overview of Available Test Security Analytics
   - John Weiner
3. Application of Response Similarity Index to Real Data
   - Amin Saiar
4. Successful Legal Application of Data Forensics
   - Jennifer Ancona Semko
The Candidate Agreement

- A Key Document
- Binding contract
- Defines the relationship
- Memorializes rights/obligations
- Makes expectations and remedies clear (if done properly)

The Candidate Agreement

- What does your Candidate Agreement say?
  - Are candidates on notice that sharing items is a breach?
  - That studying from recalled items is improper?
  - Did you reserve the right to invalidate scores? Suspend or ban access to examination? To take other action?
  - Are candidates on notice of the possible use of data forensics?
  - What are the grounds for action? Is there a "catch all"?
  - Do you regularly review your agreement language?

Security Procedures and Policies

- Do you have uniform security procedures and policies in place?
- How effectively are you communicating these points?
Can You Defend Your Actions?

- Do you have to "prove" cheating?
- Courts typically treat as contract matter:
  - Look to language of Candidate Agreement and documented policies and procedures
  - Evaluate program’s "good faith"
- General deference to exam programs (but state actors owe Constitutional protections)
- More on this in a moment . . .

Overview of Available Test Security Analytics

John Weiner

Overview of Forensic Analyses

- Credentialing Exam Context
- Use of Security Analytics
- Risks Addressed — Analytic Methods
- Example Analyses
- Practical Findings & Conclusions
Credentiaing Exam Content

- The test is used to determine professional competency
  - High stakes = motivation to "do what it takes" to pass the test

Use of Security Analytics

- Program Monitoring
- Quality Control
  - Exams, items
  - Raise issues to be investigated
  - Individual candidates, test centers, schools

- Investigation
  - Rule out problems
  - Determine scope of issues
  - Corroborative evidence

Risks Addressed

<table>
<thead>
<tr>
<th>Risk</th>
<th>Behavior</th>
<th>Analytic Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheating</td>
<td>Collusion</td>
<td>Response similarity (fixed forms)</td>
</tr>
<tr>
<td></td>
<td>Quasi-key/access</td>
<td>Testing time v. score</td>
</tr>
<tr>
<td>Piracy</td>
<td>Harvesting</td>
<td>Testing time v. Items answered</td>
</tr>
<tr>
<td>Proxy</td>
<td>Stand-in</td>
<td>n/a (biometrics, photo, proctor)</td>
</tr>
<tr>
<td>Volatile Retakes</td>
<td>Quasi-key/access</td>
<td>Test 2-Test 1 comparison (passers)</td>
</tr>
<tr>
<td>Change in</td>
<td>Compromised content</td>
<td>Pass rate analyses (item, exam)</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Council on Licensure, Enforcement and Regulation
Example: Response Similarity Analysis

Index J2 (PSI)
Regression of Max No. Matched Responses onto Raw Score

Index: Standardized residual \( Z = (Y - Y') / SD_y \)

Example: Response Similarity Analysis

Index B (Angoff)
No. Matched Incorrect Responses, Conditioned by Score Level

Index: Standardized matched incorrect \( Z = (X - M) / SD_y \)

Example: Pirating, Cheating Analysis

No. Items Answered (or Score) vs. Testing Time

Pirating

Cheating

-3, >0
+3, -3
-3 -2 -1 0 +1 +2 +3
No. Items Answered (pirating)
Score (cheating)
Example: Volatile Retake Analysis

Test Score Increase vs. Testing Time

Example: Change In Performance

Annual Exam Pass Rates,
Z test of uncorrelated proportions
*p<.01, d > 10%

Practical Findings and Conclusions

- Analytics play a useful role in test program management
- Investigating issues or events
- Raising issues to investigate
- Complement other practices
- Analytic methods are still evolving
  - Simple approaches are most practical
  - Empirical models yield reasonable base rates
- The future
  - Use of real-time analytics within the testing system will increase
  - Will the role of post hoc analytics decrease?
Application of Response Similarity Index to Real Data

Amin Saiar

Purpose and Practical Implications

- What is the purpose of response similarity indices?
  - To identify candidates with an unexpectedly similar or matching responses to another candidate
- How is an outlier determined?
  - When a candidate whose number of correct responses that match another’s is significantly different (higher) from the distribution of matched scores
- What does it mean when an outlier is found?
  - The paired test candidates may have colluded or otherwise shared information in preparation for the test

Procedure of PSI’s J2 Index

1. Compute Max_nMatch by form
   - Each candidate’s maximum # of matching responses with any other candidate
2. Conduct regression analysis
   - Test score onto Max_nMatch
3. Compute residuals
   - Observed minus predicted
4. Standardize residuals within sample
   - Conversion to z-score

![Graph](https://via.placeholder.com/150)
Research Study

- The most recent full year of administration statistics was gathered for seven high-volume high-stakes tests
- PSI's J2 Index was utilized to analyze response similarity among exams
- The standard for outlier detection was set at a value of 3.0 and above for the z-score of the residuals
- Outliers were identified in each of seven programs

<table>
<thead>
<tr>
<th>Exam</th>
<th>k</th>
<th>Cut Score</th>
<th>n*</th>
<th># of forms</th>
<th># of compares</th>
<th>Mean Max N Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Estate</td>
<td>110</td>
<td>82</td>
<td>7,139</td>
<td>9</td>
<td>1,441,578</td>
<td>91.6</td>
</tr>
<tr>
<td>Emissions Tech.</td>
<td>100</td>
<td>69</td>
<td>3,996</td>
<td>1</td>
<td>3,264,548</td>
<td>72.7</td>
</tr>
<tr>
<td>Nurse’s Aide</td>
<td>60</td>
<td>45</td>
<td>13,774</td>
<td>5</td>
<td>8,676,732</td>
<td>48.7</td>
</tr>
<tr>
<td>Personal Trainer</td>
<td>100</td>
<td>58</td>
<td>13,928</td>
<td>10</td>
<td>36,716,814</td>
<td>77.1</td>
</tr>
<tr>
<td>Electrician</td>
<td>100</td>
<td>69</td>
<td>11,357</td>
<td>3</td>
<td>9,523,862</td>
<td>65.4</td>
</tr>
<tr>
<td>Manicurist (Viet) 1</td>
<td>100</td>
<td>70</td>
<td>9,611</td>
<td>1</td>
<td>21,849,190</td>
<td>44.1</td>
</tr>
<tr>
<td>Manicurist (Viet) 2</td>
<td>90</td>
<td>63</td>
<td>6,094</td>
<td>1</td>
<td>9,887,174</td>
<td>40.7</td>
</tr>
</tbody>
</table>

*most recent full year of administration
Residuals – Real Estate

Regression Line – Emissions Tech

Residuals – Emissions Tech
### Findings

<table>
<thead>
<tr>
<th>Exam</th>
<th>k</th>
<th>n</th>
<th>% of n who are outliers</th>
<th>3x SD residual of maxnmatch</th>
<th>3x SD residual of maxnmatch / A total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Estate</td>
<td>110</td>
<td>7,139</td>
<td>0.0029</td>
<td>4.74</td>
<td>3.8%</td>
</tr>
<tr>
<td>Emissions Tech.</td>
<td>100</td>
<td>1,196</td>
<td>0.0032</td>
<td>6.15</td>
<td>6.2%</td>
</tr>
<tr>
<td>Nurse's Aide</td>
<td>60</td>
<td>13,774</td>
<td>0.0099</td>
<td>3.48</td>
<td>5.8%</td>
</tr>
<tr>
<td>Personal Trainer</td>
<td>100</td>
<td>11,928</td>
<td>0.0039</td>
<td>14.94</td>
<td>12.5%</td>
</tr>
<tr>
<td>Electrician</td>
<td>100</td>
<td>11,357</td>
<td>0.0013</td>
<td>10.77</td>
<td>10.8%</td>
</tr>
<tr>
<td>Manicurist (Viet) 1</td>
<td>100</td>
<td>9,611</td>
<td>0.0180</td>
<td>7.56</td>
<td>6.9%</td>
</tr>
<tr>
<td>Manicurist (Viet) 2</td>
<td>90</td>
<td>6,094</td>
<td>0.1198</td>
<td>7.56</td>
<td>8.4%</td>
</tr>
</tbody>
</table>

*most recent full year of admin

### Cheaters Never Prosper?

![Cheaters Never Prosper?](image)

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Max N Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

![Cheaters Never Prosper?](image)
Conclusions

- Use of J2 Index on real data suggests that it is indeed beneficial when:
  - Scanning exam programs for potential outliers in the number of response matches to identify programs that may warrant additional scrutiny
  - Corroborating reports of strange behavior at test site by determining whether a given candidate’s number of response matches is anomalous

Other Thoughts

- Some other anomalous data can be identified by J2, such as:
  - Candidates with unexpectedly dissimilar number of matched responses
    - No readily-identifiable explanation for this
    - No strong impetus to follow up on this
  - Candidates who likely cheated but failed anyway
    - No strong impetus to follow up on specific candidates who cheat and fail, due to decision outcome
    - However, this may suggest a need to monitor for potential breaches of security

Next Steps

- J2 Index and similar indices are sample-dependant
  - Cheating behaviors can impact the regression model and may distort the standard used to identify outliers
  - However, a lot of cheating would likely have to occur to make a sizable impact
- Our previous attempts at applying identification standards found in one program against another have not been fruitful (a case of apples and oranges – pardon the pun)
- We are working on identifying standards for detection that can apply to a general population of test candidates, instead of a norm-referenced standard for a given sample
Successful Legal Application of Data Forensics

Jennifer Ancona Semko

Deference... within limits

"It's not the same, I was caught stealing office supplies. You, on the other hand, got caught stealing ideas."

Example: Murray v. ETS

*Murray v. ETS, 170 F.3d 514 (5th Cir. 1999)* (SAT Exam)
- Louisiana basketball player
- Needed 820 on SAT; scored 700
- Scored 1300 on second attempt
- Similarity to nearby student (3 in 100 million odds)
- Scored 800 when offered re-take
Example: Murray v. ETS

• “ETS’s contract with Murray clearly and explicitly reserved to ETS the right to withhold any scores ETS had reason to believe were not valid. The only contractual duty ETS owed to Murray was to investigate the validity of Murray’s scores in good faith.”

Example: Langston v. ACT

Langston v. ACT, 890 F.2d 380 (11th Cir. 1989) (ACT Exam)
• Alabama football player
• Scored 10 on ACT, then 20
• Inconsistent with GPA; unusual similarity to nearby student
• “Under the governing law, the outcome of plaintiff’s case does not turn on whether or not plaintiff cheated on his exam, but only on whether or not ACT carried out its contractual obligations in good faith.”
Presenting Forensic Evidence

- The numbers can be staggering
- Example: Analysis of performances by candidates in the Philippines on health field licensure exam
- Odds of Candidate X performance occurring by chance?

One in 21,000,000,000,000,000,000,000,000

Presenting Forensic Evidence

- Forensic analysis may be useful in deterring a lawsuit
- Expert testimony can support invalidation decision or other sanction
- Prepare to defend analysis and expert’s qualifications during litigation
- Remember your audience

Presenting Forensic Evidence

- Forensic evidence and testimony of expert must satisfy Rule 702 of the Federal Rules of Evidence:
  - (1) based upon scientific facts or data;
  - (2) product of reliable principles and methods; AND
  - (3) principles and methods applied to the facts of the case.
Food For Thought

- What would trigger candidate action in your organization?
- How will you demonstrate “good faith”?
- Courts typically treat as a contract matter – thus, the Candidate Agreement is key

Speaker Contact Information

- Jennifer Ancona Semko
  - Partner, Baker & McKenzie
  - Jennifer.Semko@bakermckenzie.com
- John Weiner
  - Chief Science Officer, PSI
  - John@psionline.com
- Amin Saiar, PhD
  - Senior Psychometrician, PSI
  - Amin@psionline.com