Psychometricians Playing Detective:

Using Data Forensics Techniques for Spotting Signs of Misconduct in Test-Takers' Item Response Data

Greg Hurtz, Ph.D.
Professor
Industrial-Organizational & Quantitative Psychology
California State University, Sacramento
-and-
Senior Research Scientist
Scoring & Analytics
PSI Services LLC

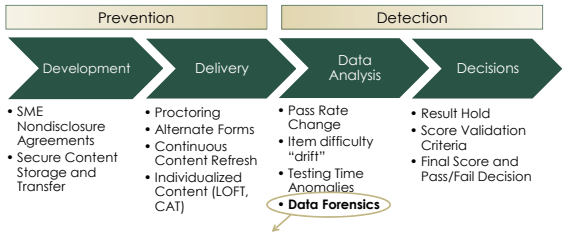
PTC-NC Summer Conference
June 27, 2019, Sacramento, CA

Example Fraudulent Test-Taker Behaviors

Traces Left in Data:	DURING	Statistical Detection:
Excessive response (A,B,C,D) similarity	• Copy answers from an adjacent test-taker. • Recall previously memorized test content. • Enter previously memorized answers. • Memorize ("harvest") new test content. • Communicate with a colluding party. • Allow proxy tester to take exam.	Matching responses
Item score (0,1) pattern irregularities		Matching errors
Item response time pattern anomalies		Mismatch of item scores with item difficulties
Response order (vs delivery order)		Short overall test time
Answer change patterns		Fast item response times
		Relations between item times and other anomalies

With 100 items, that's 100 observations of each metric (response, score, time)

Security Across the Test Development Cycle



- **Data forensics**: Application of statistical detection methods to help recognize anomalies in data patterns that are consistent with fraudulent test-taking behavior.

Introduction to Statistical Detection Methods

- We'll discuss:
  - Three categories of indices
  - One example from each category that is relatively straightforward to explain and compute
    - **Similarities:** EEIC/D
    - **Item Score Patterns:** Personal point-biserial
    - **Item Response Times:** Natural log deviations

**See research literature for other indices and their relative strengths/weaknesses; those covered here are intended to be introductory for explanatory purposes, and not overly complex statistically. We're using similar, but more refined and powerful, measures at PSI.*

Example Fraudulent Test-Taker Behaviors

BEFORE	DURING	AFTER
• Gain prior access to test content. • Solicit "inside" assistance (e.g., from a proctor). • Arrange services of a proxy tester. • Prepare for use of "spy" technology. • Arrange for a colluding party to assist during test.	• Copy answers from an adjacent test-taker. • Recall previously memorized test content. • Enter previously memorized answers. • Memorize ("harvest") new test content. • Communicate with a colluding party. • Allow proxy tester to take exam.	• Have answers changed. • Type questions from memory to share. • Share content with fellow trainees who have yet to test. • Upload questions to a "brain dump" website. • Offer fraudulent test preparation services to others. • Offer proxy testing services to others.

Similarities: EEIC/D

- **Description:**
  - **EEIC** = "**E**xact **E**rrors **I**n **C**ommon"
  - **D** = "**D**ifferences"
  - *For items that two candidates both got wrong, how often did they choose the same wrong answer, relative to the number of different responses they gave?*

Harpp, D. N., Hogan, J. J., & Jennings, J. S. (1996). Crime in the classroom: Part II. An update. *Journal of Chemical Education, 73,* 349-351.

Similarities: EEIC/D

- **Computation:**
  1. Count EEIC
  2. Count D
  3. Divide!

  – Criterion: Harpp et al. suggested values above 1 are highly suspect, especially under their study conditions. We have tended to be more conservative and use a criterion of 3.

  – Caveat: Undefined if D=0. In such cases, setting D=(1/3) ensures that two people with identical answers will reach 3, as long as they made at least one error.

Item Score Patterns: Personal point-biserial

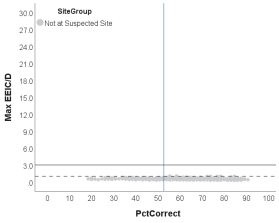
- **Computation:**
  1. Arrange data with each person's score on each item alongside each item's *p*-value
  2. Use Pearson *r* formula to compute the point-biserial correlation between the two sets of values for each person

  – Criterion: None specified.

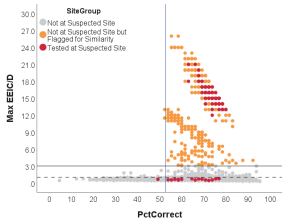
Similarities: EEIC/D

- **Examples:**

Baseline sites; EEIC/D rarely exceeds 1, never exceeds 3



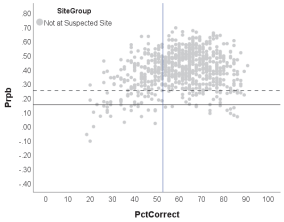
Operational run; cases from two suspected sites are flagged red. Other problematic cases are flagged orange.



Item Score Patterns: Personal point-biserial

- **Examples:**

Baseline sites; Most values are ≥ .25 or so



Operational run; cases from two suspected sites are flagged red. Cases flagged previously for similarity are flagged orange.

Item Score Patterns: Personal point-biserial

- **Description:**
  – For each test-taker this index is the correlation between their pattern of (0,1) scores and the *p*-values of the items.

  – In general, people should tend to get relatively easier items right and harder items wrong. This is a fundamental premise of CTT & IRT measurement models.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement, 28,*105–113.
(*Note*: They actually proposed correcting the correlation to get biserial *r*)

Item Response Times: Natural log deviations

- **Description:**
  – Computes each test-taker's speed as a comparison of their time spent on the items relative to items' typical "time demand".

  – More time than typical is slowness; less time is speed. Variability is expected; extreme speed may be suspect.

  – The model residual is a general measure of response time pattern irregularity.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement, 23,*195–210.

Item Response Times: Natural log deviations

- **Computation:**
  1. Compute log of seconds spent by each person on each item (transforms skew to ~normal)
  2. For each test-taker compute deviation of mean item log-time from grand mean
  3. Reflect value so it more easily represents speed (rather than slowness)
     - (*Note*: Reflecting the value was a modification in a later publication, where a more refined formula was also developed.)

  – Criterion: None specified.

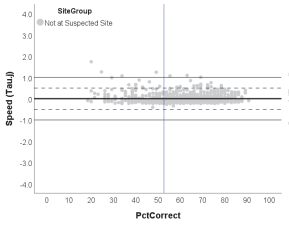
Summary and Concluding Remarks

- Item responses and response times are measures of test-taker behavior
- There is usually some regularity to this behavior, within normal ranges
- Statistical methods can be used to recognize irregularities
- Patterns and combinations of irregularities may be indicative of specific behaviors
- Statistical patterns should be corroborated with other evidence (e.g., video, proctor logs)
- These indices can be powerful tools for recognizing cheating, and potentially compromised test content.

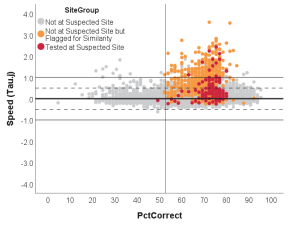
Item Response Times: Natural log deviations

- **Examples:**

Baseline sites; Most values are ≤~1



Operational run; cases from two suspected sites are flagged red. Cases flagged previously for similarity are flagged orange.



Acknowledgments

Much of this presentation is based on research and development work performed in conjunction with the following individuals at PSI Services LLC:

John Weiner, Chief Science Officer

Nicole Tucker, Director, Scoring & Analytics

Zuru Du, Senior Psychometrician

Questions?

Contact: ghurtz@csus.edu

The Detective Work: What Does it Mean?

- What behavior is likely indicated by the patterns in the data? Some example thoughts…
  – **Rapid guessing:** Probably irregular item scores, high speed, low similarity
  – **Proxy testing:** Probably high speed and similarity
    - If they try and throw you off by not going too fast, they still might have irregular time patterns
    - If they try and throw you off by picking some different wrong answers for different testers, they still might have high score irregularity
  – **Item harvesting:** Probably low speed with irregular time patterns