Detecting Answer Copying on Multiple-Choice Tests with Dichotomous Item Scores

Cengiz Zopluoglu
University of Miami

Troy T. Chen
Chi-Yu Huang
Andrew Mroch
ACT, Inc.
**Answer Copying Indices:**

- Person-fit indices vs. Answer similarity indices

- Source of Evidence:
  - Identical incorrect responses
  - Identical correct and incorrect responses
  - All items

- Type of Statistical Distribution
  - Empirical Null distribution
  - Binomial Distribution
  - Poisson Distribution
  - Compound Binomial Distribution
  - Normal Distribution
Research Purpose

– To investigate the statistical performance of answer copying indices under different simulated conditions by using dichotomous item scores
Independent Variables:

- **IRT Model**: 2PL and 3PL
- **Test Length**: 30-item and 50-item
- **Ability Group of Pairs**: Low-Low, Low-Medium, Low-High, Medium-Medium, Medium-High, High-High
- **Amount of Copying**: 20%, 40%, 60%
- **Type of Copying**: Random, Random-String

\[2 \times 2 \times 6 \times 3 \times 2 = 144\] simulated conditions for power analysis

\[2 \times 2 = 4\] conditions for Type I error rate analysis
Data Generation for One Replication

- Item parameters
- Ability parameters

Simulate Response Data for 3000 Examinees

Choose response vectors for copiers
Choose corresponding response vectors for sources

Simulate Answer Copying

Match response vectors (honest pairs of examinees)

Compute Indices Using Form B test Performance

Compute Answer Copying Indices
Statistical indices included in the study:

a. **Person-fit indices**: Lz and Modified Caution Index

b. **Answer Similarity Indices**: EMRA1, EMRA2, GBT, K and its variants (K1, K2, S1, S2), and $\omega$
Analysis:

**Power:** How many pairs are truly detected out of 5,000 simulated answer copying pairs within each condition by each index at nominal alpha level of .01?

**Type I Error Rate:** How many pairs are falsely detected out of 180,000 simulated honest pairs within each condition by each index at nominal alpha level of .01?
RESULTS
Empirical Type I Error Rates at $\alpha = .01$
Empirical Power at $\alpha = .01$
## Eta-Squared Effect Sizes from ANOVA on Statistical Power

<table>
<thead>
<tr>
<th></th>
<th>EMRA1</th>
<th>EMRA2</th>
<th>GBT</th>
<th>K</th>
<th>K1</th>
<th>K2</th>
<th>S1</th>
<th>S2</th>
<th>w</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>0.001***</td>
<td>0.001***</td>
<td>0.001***</td>
<td>0.001***</td>
<td>0.024***</td>
<td>0.024***</td>
<td><strong>0.056</strong>*</td>
<td><strong>0.050</strong>*</td>
<td>0.003***</td>
</tr>
<tr>
<td>L</td>
<td>0.091***</td>
<td><strong>0.181</strong>*</td>
<td><strong>0.094</strong>*</td>
<td><strong>0.116</strong>*</td>
<td>0.046***</td>
<td>0.030***</td>
<td>0.013***</td>
<td><strong>0.082</strong>*</td>
<td><strong>0.082</strong>*</td>
</tr>
<tr>
<td>A</td>
<td><strong>0.818</strong>*</td>
<td><strong>0.568</strong>*</td>
<td><strong>0.755</strong>*</td>
<td><strong>0.726</strong>*</td>
<td><strong>0.747</strong>*</td>
<td><strong>0.777</strong>*</td>
<td><strong>0.610</strong>*</td>
<td><strong>0.588</strong></td>
<td><strong>0.812</strong>*</td>
</tr>
<tr>
<td>T</td>
<td>&lt;.001***</td>
<td>0.001***</td>
<td>&lt;.001***</td>
<td>0.001***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.001***</td>
<td>&lt;.001**</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>G</td>
<td>0.032***</td>
<td>0.037***</td>
<td>0.036***</td>
<td>0.014***</td>
<td>0.024***</td>
<td>0.019***</td>
<td><strong>0.096</strong>*</td>
<td><strong>0.137</strong>*</td>
<td><strong>0.051</strong>*</td>
</tr>
</tbody>
</table>

- **M**: IRT Model
- **L**: Test Length
- **A**: Amount of Copying
- **T**: Type of Copying
- **G**: Ability Group
Main effect of Amount of Copying on Empirical Power at $\alpha = .01$
Main effect of Test Length on Empirical Power at $\alpha = .01$
Conclusions:

When dichotomous IRT models and dichotomous response outcomes are used:

1. The $\omega$ index showed highest detection rates, and EMRA1 and GBT also provided reasonable detection rates.

2. The K index and its variants (K1, K2, S1, S2) and EMRA 2 showed relatively lower detection rates.

3. Person-fit indices show very low power for detecting answer copying.
What's Next Step?
Thank you!

Special Thanks to:

• Tami Hrasky
• Deborah Harris
• Karen Zimmermann
• Tianli Li