Detecting Test Tampering at the Group Level

James A. Wollack
University of Wisconsin

Carol Eckerly
University of Wisconsin
• Few approaches to detection at group-level
• Unusually large score gains
  • Likely to lose power as group expands to include non-tampered individuals/classes/schools
• Empirical estimates of number of erasures
  • No clear understanding of error rates
  • No accurate probabilistic statement of the likelihood of results
• Very little is known about how well these approaches actually work
• Current study focused on a model-based approach to detect tampering at the group-level
Erasure Detection Index (EDI)

- EDI (Wollack, Cohen, & Eckerly, 2013) compares individual’s WTR score with that person’s expected WTR score
  - Expected number is estimated as the expected number correct score across all erased items
  - Appropriate IRT model is used to estimate $P(x_{ij} = 1)$
  - Estimate $\theta_j$ across non-erased items only: $\theta \downarrow j [i \notin I \downarrow E, j ]$

$$EDI = X \downarrow j, I \downarrow E, j - \sum_{i \in I \downarrow E, j} \uparrow \subseteq P(x \downarrow ij = 1) - \frac{1}{2} / \sqrt{\sum_{i \in I \downarrow E, j} \uparrow \subseteq P(x \downarrow ij = 1) [1 - P(x \downarrow ij = 1)]} = WTR - E(WTR | I \downarrow E, j) - 1/2 / SE(WTR)$$
EDI Properties

• Properties were examined in simulation study
  • Multiple types of tampering and benign erasures
  • Manipulated the ability-level of tampered student
  • 5 – 15 tampered items per student
• EDI had strong Type I error control and power
## Power of EDI for Individuals

### 5 Tampered items

<table>
<thead>
<tr>
<th>Quintile</th>
<th>.00001</th>
<th>.0001</th>
<th>.0005</th>
<th>.001</th>
<th>.005</th>
<th>.01</th>
<th>.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.140</td>
<td>.258</td>
<td>.385</td>
<td>.458</td>
<td>.676</td>
<td>.765</td>
<td>.961</td>
</tr>
<tr>
<td>2</td>
<td>.005</td>
<td>.018</td>
<td>.046</td>
<td>.075</td>
<td>.287</td>
<td>.420</td>
<td>.794</td>
</tr>
<tr>
<td>3</td>
<td>.000</td>
<td>.001</td>
<td>.007</td>
<td>.014</td>
<td>.081</td>
<td>.162</td>
<td>.605</td>
</tr>
<tr>
<td>4</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.011</td>
<td>.035</td>
<td>.304</td>
</tr>
<tr>
<td>5</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.086</td>
</tr>
</tbody>
</table>

### 10 Tampered items

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.587</td>
<td>.779</td>
<td>.888</td>
<td>.927</td>
<td>.980</td>
<td>.991</td>
<td>.999</td>
</tr>
<tr>
<td>2</td>
<td>.077</td>
<td>.250</td>
<td>.473</td>
<td>.584</td>
<td>.834</td>
<td>.904</td>
<td>.990</td>
</tr>
</tbody>
</table>
Extension of EDI to the Group Level

• Computation of EDI at student-level involves three components: WTR, E(WTR)*, and SE(WTR)*
  
  * denotes that $\theta \downarrow j[i \notin I \downarrow E, j]$ is used in place of $\theta \downarrow j$.

  $$EDI_{g} = \sum_{g} \left[ \sum_{i \in I \downarrow E, j} \frac{1}{2} \sqrt{\sum_{j} \left[ \sum_{i \in I \downarrow E, j} P(x \downarrow ij = 1) \right]} - \frac{1}{\sqrt{\sum_{j} \left[ \sum_{i \in I \downarrow E, j} P(x \downarrow ij = 1) \right]}} \right]$$

• Compute EDI components for each student in group

• Essentially treats the class as a single student taking one really long test, except that each student’s $\theta \downarrow j[i \notin I \downarrow E, j]$, erased items, and WTR data are used for summary statistic.
Simulating Erasures

• Data simulated under the nominal response model
  • 50-item test
• Included both fraudulent and benign erasures
• Within each level of fraudulent erasures studied, benign erasures were simulated for all examinees.
  • Misalignment Erasures for random 2% of examinees
    • # Misaligned ~ Bin(50, .25)
  • Random Erasures remaining 98% examinees
    • # Random erasures ~ Bin (50, .02)
    • Approximately 1/3 students had no benign erasures
Simulating Fraudulent Erasures

• Simulated on top of benign erasures
  • 1,000 replications (Schools) per condition
  • School-Level Variables
    • School Selection: Random or Mean Ability-Weighted
    • Classes/School (1, 3, 6) × % Tampered Classes (0%, 33%, 67%, 100%)
      • 0% provided null data for Type I error study
      • 33% and 67% conditions not possible with 1 Class—7 power conditions
  • Class-Level Variables
    • # Erasure Victims per class: 1, 3, 5, 10
    • Victim Selection: Random or Ability-Weighted
    • # Tampered Items per victim: 3, 5, 10
    • Class size: 15, 25, 35
  • Tampered questions were simulated to be answered correctly
• α (7 levels): .05, .01, .005, .001, .0005, .0001, .00001
Implementation and Evaluation

- Nominal response model used to estimate $P(x_{ij} = 1)$
  - Could have also used a dichotomous model
- Item parameters treated as known
  - No attempt was made to mirror reality with respect to amounts and magnitudes of tampering
- EDI computed
  - At Individual Student Level
  - At Class Level
  - At School Level
- Evaluative Measures
  - Type I Error rate and Power at each of the three levels
  - Only results from Random School Selection are presented
    - Class and School-Level only
Type I error results

Over all null conditions

<table>
<thead>
<tr>
<th>Level</th>
<th>.00001</th>
<th>.0001</th>
<th>.0005</th>
<th>.001</th>
<th>.005</th>
<th>.01</th>
<th>.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>0.00000</td>
<td>0.0000</td>
<td>0.0002</td>
<td>0.0004</td>
<td>0.0022</td>
<td>0.005</td>
<td>0.029</td>
</tr>
<tr>
<td>School</td>
<td>0.00000</td>
<td>0.0001</td>
<td>0.0003</td>
<td>0.0006</td>
<td>0.0035</td>
<td>0.007</td>
<td>0.037</td>
</tr>
</tbody>
</table>
Class-Level Power

Three Erased Items

Five Erased Items

Ten Erased Items

5 Erased, 5 Victims

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>0.99</td>
</tr>
<tr>
<td>.01</td>
<td>0.94</td>
</tr>
<tr>
<td>.005</td>
<td>0.91</td>
</tr>
<tr>
<td>.001</td>
<td>0.82</td>
</tr>
<tr>
<td>.0005</td>
<td>0.77</td>
</tr>
<tr>
<td>.0001</td>
<td>0.66</td>
</tr>
<tr>
<td>.00001</td>
<td>0.49</td>
</tr>
</tbody>
</table>
School-Level Power: 3 Erased Items

- 1 Victim
- 3 Victims
- 5 Victims
- 10 Victims

Power

# Tampered Classes

α

0.05
0.01
0.005
0.001
0.0005
0.00001
School-Level Power: 5 Erased Items

- 1 Victim
- 3 Victims
- 5 Victims
- 10 Victims

Power

# Tampered Classes

α

0.05
0.01
0.005
0.001
0.0005
0.0001
School-Level Power: 10 Erased Items

![Graph showing the relationship between the number of tampered classes and power for different numbers of victims. The graph includes lines for different significance levels (α) ranging from 0.05 to 0.00001.](image)

- **Power** is plotted on the y-axis, ranging from 0 to 1.0.
- **# Tampered Classes** is plotted on the x-axis, ranging from 1 to 6.
- The graph is divided into sections for 1, 3, 5, and 10 victims, each with different lines representing varying significance levels.
Conclusion

• EDI appears to work very well for group-level tampering detection.
  • Type I error rate was well controlled at nearly all α levels
    • Small amounts of inflation evident within high-ability schools
  • Power was quite strong, even when few items were tampered for relatively small numbers of students, and at small α levels
Thank You

For more information, contact:

James Wollack
University of Wisconsin
jwollack@wisc.edu