Measuring the Reliability of Diagnostic Mastery Classifications at Multiple Levels of Reporting

Jake Thompson, Amy Clark, & Brooke Nash
ATLAS, University of Kansas
A BRIEF OVERVIEW OF DIAGNOSTIC CLASSIFICATION MODELS
Diagnostic Classification Models

- Latent trait models that assume a categorical latent trait
- Multivariate
- Probability of a correct response determined by the examinees’ attribute profiles and a Q-matrix
- Scores are based on an examinee’s probability of mastery on the defined attributes
Reliability in DCMs

- Traditional methods are inadequate
- Templin & Bradshaw (2013)
  - Use mastery probabilities to create a 2x2 contingency table for re-test mastery
  - Aggregate over all examinees
  - Reliability estimate is the tetrachoric correlation of aggregated contingency table
- Provides a reliability estimate for each attribute
SO WHAT’S THE PROBLEM?
Using DCMs in a Learning Map Setting

• Thousands of possible nodes in the map structure
• On any given blueprint examinees test on 50-100 attributes
• Fine grained inferences, but can be overwhelming
Aggregated Attribute Summaries
Reliability of the Aggregation

1. Draw with replacement a student from the operational data set
2. Simulate new item responses based on model parameters and student mastery status
3. Score simulated item responses
4. Calculate simulated aggregations
5. Compare simulated scores to observed scores
Summarize Attribute Agreement
## Summarize Content Standard Agreement

<table>
<thead>
<tr>
<th>Metric</th>
<th>Index range</th>
<th>&lt;.60</th>
<th>.60–.64</th>
<th>.65–.69</th>
<th>.70–.74</th>
<th>.75–.79</th>
<th>.80–.84</th>
<th>.85–.89</th>
<th>.90–.94</th>
<th>.95–1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polychoric correlation</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>32</td>
<td>81</td>
<td>20</td>
</tr>
<tr>
<td>Correct classification rate</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>16</td>
<td>58</td>
<td>57</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Cohen’s kappa</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>20</td>
<td>59</td>
<td>52</td>
<td>5</td>
</tr>
</tbody>
</table>
## Summarize Subject Agreement

<table>
<thead>
<tr>
<th>Grade</th>
<th>Skills mastered correlation</th>
<th>Average student correct classification</th>
<th>Average student Cohen’s kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.981</td>
<td>.982</td>
<td>.963</td>
</tr>
<tr>
<td>4</td>
<td>.983</td>
<td>.984</td>
<td>.966</td>
</tr>
<tr>
<td>5</td>
<td>.979</td>
<td>.978</td>
<td>.952</td>
</tr>
<tr>
<td>6</td>
<td>.976</td>
<td>.974</td>
<td>.943</td>
</tr>
<tr>
<td>7</td>
<td>.964</td>
<td>.965</td>
<td>.919</td>
</tr>
<tr>
<td>8</td>
<td>.971</td>
<td>.968</td>
<td>.927</td>
</tr>
<tr>
<td>9</td>
<td>.980</td>
<td>.977</td>
<td>.948</td>
</tr>
<tr>
<td>10</td>
<td>.980</td>
<td>.977</td>
<td>.947</td>
</tr>
<tr>
<td>11</td>
<td>.974</td>
<td>.967</td>
<td>.923</td>
</tr>
<tr>
<td>12</td>
<td>.969</td>
<td>.985</td>
<td>.964</td>
</tr>
</tbody>
</table>
Conclusions and Limitations

• Reporting of aggregated scores requires evidence to support the aggregates
• Simulation is one possible solution
• Limitations
  – Assumes Model fit
  – Estimates are an upper bound
  – Computationally intensive
More Information

@wjakethompson

wjakethompson@ku.edu

wjakethompson.com
dynamiclearningmaps.com