Explaining Assessment Results for Instruction Using the Cognitive Diagnostic Models

Yi Du, Ph.D.
Data Recognition Corporation
CCSSO
June 29, 2012 Minneapolis, Minnesota
Do Statewide Assessments Provide Sufficient Information for Teachers?

Based on a survey by J. P. Leighton, M. J. Gierl

- 51% of teachers believed that the state mandated assessments didn’t provide sufficient amount of information regarding students’ strengths and weakness
- 53% of teachers held the same believes about large-scale commercial assessments
- 71% of teachers believed that it would be more valuable to have more diagnostic information.

2012 Data Recognition Corporation
What Kinds of Diagnostic Information Teachers Are Specifically Interested In?

- Special skills/knowledge individual students demonstrated
- Special skills/knowledge individual students need to develop
- Strategies individual students might use to improve their skills/knowledge
- Strategies teachers might use to address student needs
What Psychometric Measurement Models (e.g., IRT) Do for Assessments?

- Approximate a person’s location on a underlying variable of interest, e.g., achievement or aptitude
- Measure unobserved “abilities” based on observed responses
- Link across multiple test forms and “adjust” for difficulty differences of tests
- Compare item difficulty and examinee ability on the same scale
What Psychometric Measurement Models Didn’t Do?

- Incorporate substantive psychological theory to explain IRT
- Realistic assumptions about the psychological dependencies and variables influencing test item performance
- Explicit delineation of the psychological process that collectively reflect the construct measures by a test
- Meet the increasing demand for information about students’ cognitive processing
Extension of IRT: Cognitive Diagnostic Models

- A great potential for innovation and change in today’s demanding of information
- Integration of educational measurement and cognitive psychology
- Design tests targeted for information about students' cognitive strengths and weaknesses
- Maintain merits of IRT and increasing diagnostic functions
The General Cognitive Diagnostic Model

Assuming a simple form of \( h(q,a) \) and introduce an additional restriction on \( \gamma \).

\[
\gamma_{xik} = x\gamma_{ik} \text{ and } h(q_{ik}, a_{k}) = q_{ik} a_{k}
\]

Which yields the model

\[
p(x \mid \beta_i, a, q_i, \gamma_i) = \frac{\exp[\beta_{xi} + \sum_{k=1}^{k} x\gamma_{ik} q_{ik} a_{k}]}{1 + \sum_{y=1}^{m} \exp[\beta_{xi} + \sum_{k=1}^{k} x\gamma_{ik} q_{ik} a_{k}]}
\]
Cognitive Diagnosis Models (GDM)

- Providing a tool to extract the needed information – these underlying latent traits
- These latent traits can be skills, knowledge, strategies, cognitive processing, etc.
- GDM models might provide diagnostic information from assessments for instruction
- The software MDLTD (von Davier, 2005) makes this approach feasible
What Are Similarities and Differences Between IRT and Cognitive Diagnostic Models

- Cognitive Diagnostic Models, like IRT, related item responses to an unobserved variable (say testing strategy type)
- Response probabilities are class specific or skill specific, depending upon models
- Item responses are independent within classes (or skills)
- Each item has class dependent (or skill dependent) for a response vector assume local independence
Brief History of the General Diagnostic Models

- Multivariate IRT (Reckase, …), or Multivariate Rasch Models (Adams et al…)
- Multiple Classification Latent Class Models (Haertel, Maris, …)
- Rule Space (Tatsuoka), Fusion-Arpeggio (stout et al.), Bayes Network (Mislevy, Almond, …)
- Mixture IRT Models (Mislevy & Verhelst, Rost, von Davier, Yamamoto, …)
A Study on Cognitive Diagnosis for Statewide Assessments

- This study explores the possibility of using the cognitive diagnostic models (GDM) into statewide assessments

- Testing data from a statewide grade 11 ELA testing data were explored via GDM to examine:
  - What attributes (or skills) were assessed
  - How many levels of the attributes are appropriate to report
  - What students’ performance on different levels of attributes
  - What are different class profiles?
Data

- Data for this study was from a state grade 11 English Language Arts spring operational assessment.
- The assessment includes one writing prompt, constructed response, and multiple-choice items.
- Data were obtained by randomly selecting 5000 students from the population (40,000).
- Assessment was constructed based on unidimensional assumption.
The English Language Assessment

- Exploring three possible skills (dimensions or attributes) in the assessment
  - Writing skill
  - Reading skill for literacy
  - Reading skill for information usage
Analysis

- Examine the dimensionality of the assessment
- Diagnosis of which cognitive diagnosis model best fits the data
- Explore the cognitive attributes under students’ responses
- Estimate the distribution of latent attribute patterns for subgroups
- Explore the item performance by different dimension (skills)
GDM Model Fit

Different GDM models were compared to see which one is reasonably consistent with the data and so does not require re-specification.

Three criteria are used to evaluate models:

- the log-likelihood
- The Akaike Information Criterion (AIC)
- the Bayesian Information Criteria Index (BIC)
## Model Comparison for 1PL GDMs

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>Log Likelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-level-one-scale 1PL GDM</td>
<td>66</td>
<td>-153463.50</td>
<td>307059.0</td>
<td>307489.1</td>
</tr>
<tr>
<td>Three-level-one-scale 1PL GDM</td>
<td>67</td>
<td>-147761.20</td>
<td>295656.4</td>
<td>296093.0</td>
</tr>
<tr>
<td>Four-level-one-scale 1PL GDM</td>
<td>68</td>
<td>-145578.35</td>
<td>291292.7</td>
<td>291735.8</td>
</tr>
<tr>
<td>Two-level-two-scale 1PL GDM</td>
<td>68</td>
<td>-152873.09</td>
<td>305882.2</td>
<td>306325.3</td>
</tr>
<tr>
<td>Three-level-two-scale 1PL GDM</td>
<td>73</td>
<td>-147430.12</td>
<td>295006.2</td>
<td>295482.0</td>
</tr>
<tr>
<td><strong>Four-level-two-scale 1PL GDM</strong></td>
<td><strong>80</strong></td>
<td><strong>-145470.17</strong></td>
<td><strong>291100.3</strong></td>
<td><strong>291621.7</strong></td>
</tr>
<tr>
<td>Two-level-three-scale 1PL GDM</td>
<td>72</td>
<td>-153526.02</td>
<td>307196.0</td>
<td>307665.2</td>
</tr>
<tr>
<td>Three-level-three-scale 1PL GDM</td>
<td>91</td>
<td>-147992.09</td>
<td>296166.2</td>
<td>296759.2</td>
</tr>
<tr>
<td>Four-level-three-scale 1PL GDM</td>
<td>128</td>
<td>-146244.35</td>
<td>292744.7</td>
<td>293578.9</td>
</tr>
</tbody>
</table>
# Model Comparison for the 2PL GDMs

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
<th>Log Likelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-level-one-scale 2PL GDM</td>
<td>113</td>
<td>-147643.38</td>
<td>295512.76</td>
<td>296249.15</td>
</tr>
<tr>
<td>Three-level-one-scale 2PL GDM</td>
<td>114</td>
<td>-145442.38</td>
<td>291112.75</td>
<td>291855.67</td>
</tr>
<tr>
<td>Four-level-one-scale 2PL GDM</td>
<td>115</td>
<td>-144628.38</td>
<td>289486.76</td>
<td>290236.19</td>
</tr>
<tr>
<td>Two-level-two-scale 2PL GDM</td>
<td>117</td>
<td>-147326.32</td>
<td>294886.65</td>
<td>295649.11</td>
</tr>
<tr>
<td>Three-level-two-scale 2PL GDM</td>
<td>122</td>
<td>-145133.98</td>
<td>290511.95</td>
<td>291307</td>
</tr>
<tr>
<td>Four-level-two-scale 2PL GDM</td>
<td>129</td>
<td>-144293.01</td>
<td>288844.01</td>
<td>289684.68</td>
</tr>
<tr>
<td>Two-level-three-scale 2PL GDM</td>
<td>126</td>
<td>-147324.89</td>
<td>294901.79</td>
<td>295722.9</td>
</tr>
<tr>
<td>Three-level-three-scale 1PL GDM</td>
<td>145</td>
<td>-145265.15</td>
<td>290820.31</td>
<td>291765.24</td>
</tr>
<tr>
<td>Four-level-three-scale 1PL GDM</td>
<td>182</td>
<td>-144399.55</td>
<td>289163.11</td>
<td>290349.16</td>
</tr>
</tbody>
</table>
Male and Female Comparison (Theta)

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th></th>
<th>Male</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>STD</td>
<td>Mean</td>
<td>STD</td>
</tr>
<tr>
<td>Writing</td>
<td>0.79</td>
<td>0.81</td>
<td>0.60</td>
<td>0.85</td>
</tr>
<tr>
<td>Reading</td>
<td>0.59</td>
<td>0.86</td>
<td>0.50</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Writing Skill Comparison: Male and Female

Percentages of Students at Each Level

Skill Level

Female
Male

2012 Data Recognition Corporation
Reading Skill Comparison: Male and Female

Percentages of Students at Each Level

Skill Level

Female

Male
White and Black Student Comparison (Theta)

<table>
<thead>
<tr>
<th></th>
<th>Black</th>
<th></th>
<th>White</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>STD</td>
<td>Mean</td>
<td>STD</td>
</tr>
<tr>
<td>Writing</td>
<td>0.54</td>
<td>0.84</td>
<td>0.76</td>
<td>0.85</td>
</tr>
<tr>
<td>Reading</td>
<td>0.46</td>
<td>0.79</td>
<td>0.50</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Writing Skill Comparison: Black and White

Percentages of Students at Each Level

Skill Level

Black
White
Reading Skill Comparison: Black and White

Percentages of Students at Each Level

Skill Level

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7

1 2 3 4

Black

White
### Student SES Comparison (Theta)

<table>
<thead>
<tr>
<th></th>
<th>Free Lunch</th>
<th>Pain Lunch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>STD</td>
</tr>
<tr>
<td>Writing</td>
<td>-0.42</td>
<td>1.46</td>
</tr>
<tr>
<td>Reading</td>
<td>-0.77</td>
<td>1.07</td>
</tr>
</tbody>
</table>
Writing Skill Comparison: Free and Paid Lunch

Percentages of Students at Each Level

Skill Level

Free
Paid
Reading Skill Comparison: Free and Paid Lunch

Percentages of Students at Each Level

Skill Level

Free
Pain
2 Class Latent Class Profiles

Each item can have a different probability in each class, so classes have different profiles.
Results

- Two scales, writing and reading, were identified from the English language arts assessment.
- The number of the levels of the latent attribute affect the model fitting greater than model difference or the dimensionality.
- Model differences made great effect on the model fit than dimensionality.
- The level of attribute revealed more detailed information which are useful in classroom instruction.
Findings and Discussion

Cognitive Diagnostic Models can

- identify latent skills/knowledge based on data effectively
- Provide detailed information and distributions about students’ levels on these skills
- Potential to report skill profiles or group profiles to help classroom instruction and student learning, and
- More…
Findings and Discussion, Cont.

More to go on the Cognitive Diagnostic Models:

- Integrate CDM with the Evidence Center Design (ECD)
- Various statistical models
- Translation diagnostic information into teaching practices.
Possible Utility of General Diagnostic Model in Statewide Assessments

- Integrate with the Evidence-Based Design
- Identify latent variables (skills) using students’ responses
- Report student scores and the levels of latent skills based on different class profiles
- Report item performance based different class profiles
- The software MDLTD (von Davier, 2005) is user-friendly
Selected References


