Quantifying Reliability in Diagnostic Classification Models

Jonathan Templin
The University of Georgia

Robert Henson
The University of North Carolina at Greensboro
Talk Overview

• Reliability for latent variables
  - Item Response Model (IRT) Reliability
  - Diagnostic Classification Model (DCM) Reliability

• Comparing DCM reliability to IRT reliability
  - Attempt to characterize DCM estimates in an understandable way
  - Theoretical results
    - Unidimensional Rasch models
  - Simulation study
    - Varying dimensions and discrimination
  - Empirical application
    - Reliability for all models presented previously
    - Comparing unidimensional IRT and DCMs

• Implications for large scale testing
• DCMs are psychometric models with categorical latent variables
  ➢ The goal is classification
    ♦ Scaling with a finite set of points

• Quantification of reliability must reflect categorical latent variables of DCMs
  ➢ Should characterize degree of precision in latent variable
  ➢ Can also be used to compare to other latent variable methods
True Scores v. Latent Variables

• Classical notions of reliability were based on test scores
  ➢ Not latent variables directly

• Speaking of reliability in IRT, Lord (1980, p. 46):

  “True score \( \xi \) and ability \( \theta \) are the same thing expressed on different scales of measurement”

• We focus on latent variable reliability in DCMs
  ➢ Easier to compare reliability between DCMs and IRT
Classical Reliability

- Defined as a consequence of classical true score theory:
  \[ X_i = T_i + E_i \]

- \( X_i \) – observed test score; \( T_i \) – true score; \( E_i \) – error
  - \( T \) and \( E \) are independent, meaning:
    \[ \sigma^2_X = \sigma^2_T + \sigma^2_E \]

- Reliability can be considered proportion of total variance accounted for by true score:
  \[ \rho_\xi = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E} \]
IRT Reliability

• For an examinee i, mapping IRT onto CTT theory:

\[ \hat{\theta}_i = \theta_i + \epsilon_i \]

- Estimated \( \theta \)
- True \( \theta \)
- Measurement Error

• The estimated \( \theta \) consists of the true value of an examinee’s \( \theta (\theta_i) \) and measurement error \( (\epsilon_i) \)
IRT Reliability

True $\theta$

$\theta_i = \mu_\theta + \phi_i$

Population Mean

$E(\theta_i) = \mu_\theta$

Deviation from Mean

$Var(\theta_i) = \sigma^2_\theta$

Prior Distribution for $\theta$

$\theta_i \sim N(\mu_\theta, \sigma^2_\theta)$
IRT Reliability Derivation

• Placing the model for the true $\theta$ into our original equation:

$$\hat{\theta}_i = \theta_i + \varepsilon_i = \mu_\theta + \phi_i + \varepsilon_i$$

• Assuming independence of error terms and $\theta$, the total variance of the estimate is then:

$$\sigma^2_{\hat{\theta}} = \sigma^2_\theta + \sigma^2_\varepsilon$$

• Here, error variance is the inverse of test information
  - Depends on the location of $\theta$
  - Also known as the posterior variance of $\theta$
  - Andrich (1982); Briggs & Wilson (2007); Green et al. (1984); Mislevy (1993)...

\[\varepsilon_\phi \mu \varepsilon_\theta\]
From CTT to IRT Reliability

- CTT Reliability:

\[ \rho_\xi = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \]

- IRT Reliability

\[ \rho_\theta = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \overline{\sigma_\varepsilon}^2} \]

\[ \overline{\sigma_\varepsilon}^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_{\varepsilon,i}^2 \]

Average Posterior Variance for \( \theta \)
RELIABILITY IN DCMS
Latent Variables in DCMs

- Unlike $\xi$ in CTT and $\theta$ in IRT, DCMs have categorical latent variables, $\alpha_i$
  - DCMs typically use two category levels
    - Mastery/Non-mastery
    - Proficient/Not Proficient
- For a single dimension, $\alpha_i$ follows a Bernoulli Distribution
  \[ \alpha_i \sim B(p_\alpha) \]
- Here, $p_\alpha$ is the proportion of masters in the population
  - Typically estimated from data (can also be fixed)
- Therefore:
  \[ \sigma^2_\alpha = p_\alpha (1 - p_\alpha) \]
DCM Examinee Estimates

• As DCMs feature categorical latent variables, examinee estimates are reported in the form of Bernoulli variables.

• For an examinee $i$, $\hat{\alpha}_i \sim B(\hat{p}_\alpha)$
  ➢ Here $\hat{p}_\alpha$ is the (posterior) probability of mastery for the attribute.

• Therefore, the error variance associated with $\hat{\alpha}_i$ is:

$$\sigma^2_\varepsilon = \hat{p}_\alpha (1 - \hat{p}_\alpha)$$
From IRT to DCM Reliability

- IRT Reliability

\[ \rho_{\theta} = \frac{\sigma_{\theta}^2}{\sigma_{\theta}^2 + \bar{\sigma}_\varepsilon^2} \]

\[ \bar{\sigma}_\varepsilon^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_{\varepsilon,i}^2 \]

Average Posterior Variance for \( \theta \)

- DCM Reliability

\[ \rho_{\alpha} = \frac{\sigma_{\alpha}^2}{\sigma_{\alpha}^2 + \bar{\sigma}_\varepsilon^2} \]

\[ \bar{\sigma}_\varepsilon^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_{\varepsilon,i}^2 \]

Average Posterior Variance for \( \alpha \)
COMPARING DCM
AND IRT RELIABILITY
To investigate the properties of reliability in DCMs, we present results from three studies:

- **Theoretical comparisons**
  - IRT: Rasch Model; DCM: Analogous Rasch Model

- **Simulation study**
  - Generated 100 tests per condition with IRT
  - Varied number of dimensions (1 or 3), and number of items per dimension (5, 10, or 15).

- **Empirical results**
  - Large scale end-of-grade reading test
Overall Summary of Findings

• Using our metric of reliability:
  - DCM reliability was uniformly higher than IRT reliability
  - DCM reliability was never lower than IRT reliability

• Questions remain:
  - Is our metric appropriate?
  - Is it an accurate way to characterize latent variables?
  - Are comparisons valid?

• Results shown try to highlight implications of this reliability metric
Theoretical Rasch Model Reliability

<table>
<thead>
<tr>
<th>Reliability Level</th>
<th>DCM</th>
<th>IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>8 Items</td>
<td>34 Items</td>
</tr>
<tr>
<td>0.85</td>
<td>10 Items</td>
<td>48 Items</td>
</tr>
<tr>
<td>0.90</td>
<td>13 Items</td>
<td>77 Items</td>
</tr>
</tbody>
</table>
Simulation Study Results

![Graph showing the relationship between reliability and items per dimension for 1D and 3D models under DCM and IRT frameworks. The graph indicates that for both DCM and IRT, reliability increases with an increase in items per dimension. Under DCM, the reliability for 3D and 1D models is compared, with 3D generally showing higher reliability. Under IRT, a similar trend is observed, with reliability increasing as items per dimension increase.](image-url)
Empirical Results

- DCM
- IRT

- 1-Dimension
- 2-Dimension
- BiFactor

Reliability vs. Dimensional Model
Unidimensional Example

- Estimated a unidimensional DCM:
  - Two, three, four, and five categories
  - Category levels represent proficiency standards categories
    - Two categories: proficient/not proficient
    - Five categories: match state

- Constructed tests with 3-73 items
  - Best items based on estimated DCM parameters

- Calculated the reliability for $\alpha$
  - Compared to estimated 2PL reliability of 0.87 for $\theta$
Reliability

2PL $\rho_\theta = .87$

Constructing Shorter Tests

- **2 Category**: 24 Items
- **3 Category**: 42 Items
- **4 Category**: 50 Items
- **5 Category**: 54 Items

Number of Items

Reliability
CONCLUDING REMARKS
Concluding Remarks

• In this talk we:
  - Defined a reliability coefficient for DCMs
  - Showed how DCMs had higher reliability than IRT models

• Ramifications of DCM reliability:
  - Reliable measurement of multiple dimensions is possible
    - Two-attribute DCM application to empirical data:
      - Reliabilities of 0.95 and 0.90 (compared to 0.72 and 0.70 for IRT)
  - Shorter unidimensional tests
    - Two-category unidimensional DCM application to empirical data:
      - Test needed only 24 items to have same reliability as IRT with 73 items

• Questions remain about approach
  - Would other metrics be better?
Concluding Remarks

• Paradox of DCMs:
  - Sacrifice fine-grained measurement of $\theta$ for only several categories of $\alpha$
  - Increased capacity to measure ability multidimensionally

• Practical implications:
  - Multidimensional proficiency standards
    - Students must demonstrate proficiency on multiple latent attributes to be considered proficient for an overall content area
  - “Teaching to the test” would therefore represent covering more curricular content to best prepare students