Practical Adaptations of Cognitive Diagnosis Models

Jonathan Templin
Department of Psychology
University of Kansas

Robert Henson
Department of Educational Research Methodology
The University of North Carolina at Greensboro

Introduction

- Recent national education policy decisions have placed an emphasis on educational testing.
  - Used to assess the quality of instruction students are receiving.
  - No Child Left Behind has ushered in an era of “accountability.”
- Teachers are now spending large amounts of time preparing their student for standardized assessments.
- Often, preparing students for tests is at odds with the curriculum.

Introduction

- Because of this divide, there is an opportunity to change the way tests are constructed and analyzed.
  - Allowing teachers to teach to the curriculum instead of teaching strategies for taken standardized tests.
- Furthermore, there is a need to provide formative assessment to students.
  - Providing more informative paths for advancement or remediation.

Introduction

- Cognitive Diagnosis Models (CDMs) provide detailed information about the extent of knowledge a student possesses.
  - *Potentially* filling the needs of post-NCLB assessments.
- Not all CDMs are valid in all applications.
  - Currently used models can be ineffective when applied to tests created for measurement of continua.
  - Often the user will not know that the model is inappropriate.
  - Development of new models must be sensitive to realistic construction of tests.

Outline

Today's talk will feature an introduction to the concepts of cognitive diagnosis:

- Concepts underlying models for cognitive diagnosis.
- Common measurement models used for assessing skills.

Along with several examples of when cognitive diagnosis models break down:

- Problems indicative of poor analyses.
- Description of models intended to resolve issues.
- Introduction of a new model designed to compensate for deficiencies of current models.

Cognitive Diagnosis or Skills Assessment?

- Prior to discussing the concepts of cognitive diagnosis, I must note that there is some ambiguity in the name of these models.

- Exactly why these models use the term “cognitive” is unclear (although where this term came from is clear).

  - *Cognitively Diagnostic Assessment* (1995) edited by Chipman, Nichols, and Brennen.

- Such models are also called models for skills assessment, or models for skills diagnosis.

- “Skills” seems to be preferred by many due to their rooting in natural language.

Concepts

Imagine a test covering basic math:

1.) $2+3-1$ 2.) $4 \div 2$ 3.) $3 \times (4-2)$

- Using traditional assessment methods, an individual's score, or general math ability, could be estimated.
- Instead, math ability can be expressed as a set of basic skills (commonly called attributes):
  - Add
  - Subtract
  - Multiply
  - Divide
- Cognitive diagnosis models estimate a profile of the skills an individual has mastered.

Example Q-Matrix

Math Test Example Q-matrix

	Add	Sub	Mult	Div
$2+3-1$	1	1	0	0
$4/2$	0	0	0	1
$3\times(4-2)$	0	1	1	0

- Unlike unidimensional IRT, not every item measures each attribute.
- A Q-matrix indicates which attributes are measured by each item.
- Notice that the Q-matrix defines the nature of each attribute.

Example Test Items and Examinees

Math Test Example Q-matrix

	Add	Sub	Mult	Div
$2+3-1$	1	1	0	0
$4/2$	0	0	0	1
$3 \times (4-2)$	0	1	1	0

Possible Attribute Patterns

	Add	Sub	Mult	Div		Expected Correct Responses
$\boldsymbol{\alpha}_1$	1	0	0	0	$\rightarrow$	None
$\boldsymbol{\alpha}_2$	1	1	0	0	$\rightarrow$	#1
$\boldsymbol{\alpha}_3$	1	1	0	1	$\rightarrow$	#1, #2

Other Example Q-matrix

Of course, from our Napoleon Dynamite clip, the Q-matrix would look like:

	Nunchucks	Bowhunting	Computer Hacking
Girls	1	1	1

Possible Attribute Patterns

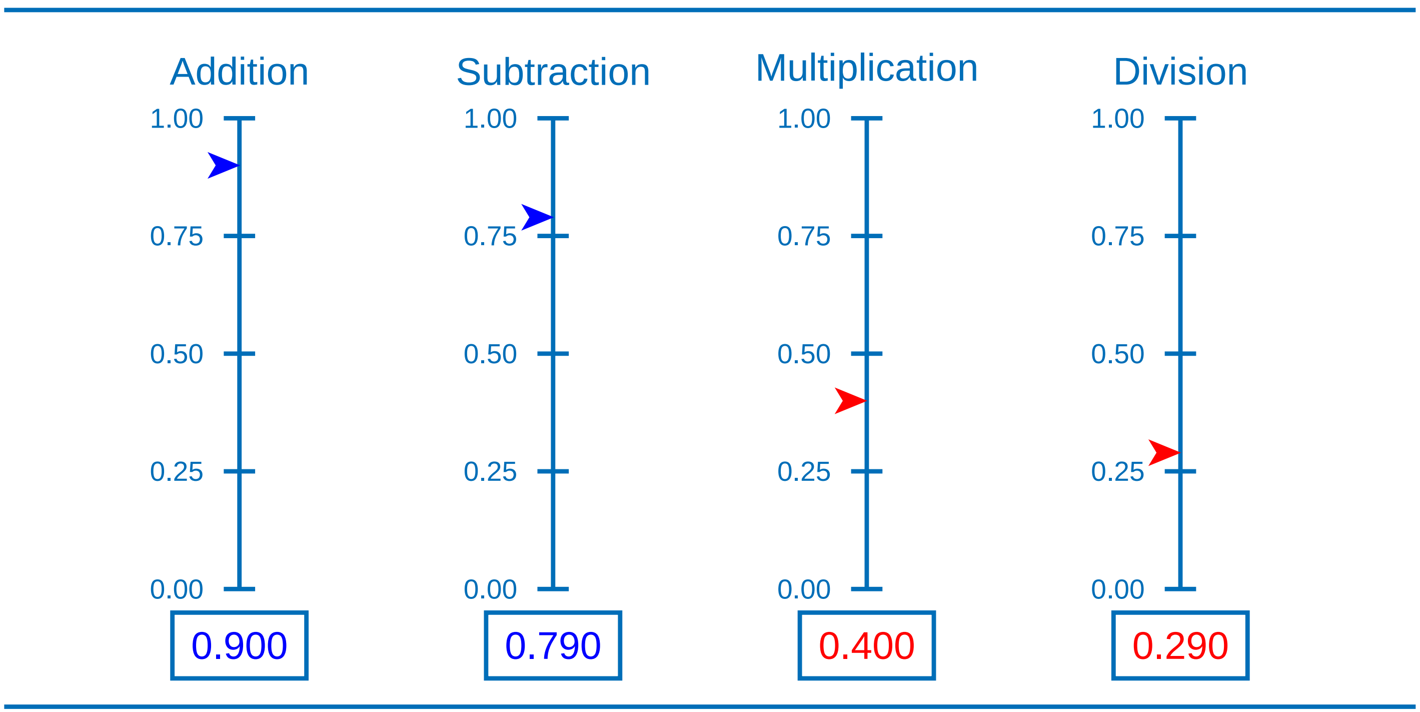
	N	B	CH		Expected Response
Napoleon	0	0	0	$\longrightarrow$	None

Cognitive Diagnosis Models...

- Are latent class models with a set equality constraints placed on class probabilities.
  - Latent classes are defined by a set of dichotomous attributes.
- Provide **why** students are not performing well in addition to **which** individuals are not performing well.
- Such information comes in the form of a posterior probability of mastery for each skill measured by the Q-matrix.

Examinee Skill Assessment Estimates

Posterior probabilities of attribute mastery:



Color Key:

➤ Probable Master

➤ Probable Non-master

Reading Skills

- We introduce a data set...
- Imagine we wanted to determine groups of examinees based on their performance on this test.
  - We could do a standard Latent Class Analysis.
  - We could an analysis with a CDM.

Reading Skills

- Data come from a large standardized test.
  - A total of 38 items.
  - Sample of 2,952 examinees.
- CDM Q-matrix has four attributes.
  - Average 1.32 attributes per item.
- Test originally developed to measure reading ability on a single latent continuum.
- Imagine we wanted to determine groups of examinees based on their performance on this test.
  - We could do a standard Latent Class Analysis.
  - We could an analysis with a CDM.

Latent Class Analysis

- Latent class models are commonly attributed to Lazarsfeld and Henry (1968).
- Use of latent class models in educational measurement date to Macready and Dayton (1977).
- The final number of latent classes is not usually predetermined prior to analysis with latent class models.
    - But is determined through comparison of posterior fit statistics.
    - This is unlike CDMs.

Latent Class Analysis

A latent class model for the response vector of individual i with $c = 1, \ldots, C$ classes:

$$P(\mathbf{X}_i = \mathbf{x}_i) = \sum_{c=1}^{C} \overset{\text{Structural}}{\overbrace{P(c)}} \prod_{j=1}^{J} \overset{\text{Measurement}}{\overbrace{P(X_{ij} = 1|c)}}$$

- $P(c)$ is the probability that any individual is a member of class c (must sum to one).
  - $c - 1$ parameters.
- $P(X_{ij} = 1|c)$ is the probability individual i answers item j correctly, given that individual i is a member of class c.
  - Part where specific CDMs are defined.

How Many Classes?

- In a latent class analysis, the first question asked is "how many classes are needed to describe my data?"
- The answer to this comes from fitting models with increasing numbers of classes and examining the relative fit of the models.
- An index of fit sometimes used is the BIC (lowest is best):

Classes	BIC
1	130,990.893
2	117,824.327
3	115,401.287
4	115,072.590
5	115,152.977

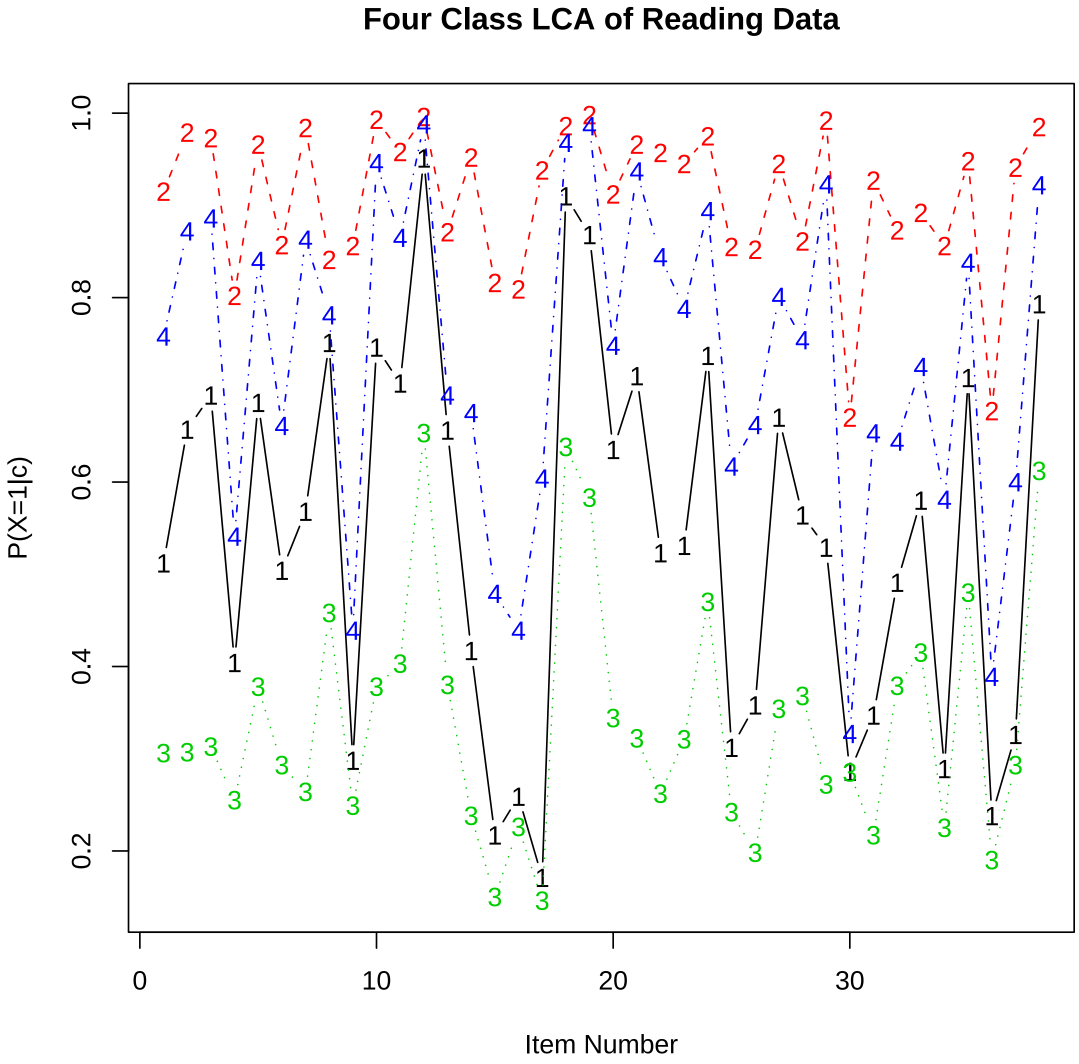
- For this application, the four-class model is considered the best fitting by the BIC.

Four Class Solution

Estimated Class Membership Probabilities:

c	$P(c)$
1	0.263
2	0.255
3	0.134
4	0.348

Estimated Item Response Probabilities:



LCA Local Independence

- LCA has the property of local independence - that given class, item responses are independent.
- To give an example, consider Items 1 and 2:

	$P(X_j = 1\|c)$			
Item j	$c = 1$	$c = 2$	$c = 3$	$c = 4$
1	0.512	0.915	0.306	0.758
2	0.657	0.979	0.307	0.872

	Item 1		
Item 2	0	1	Marginal
0	0.164	0.139	0.343
1	0.321	0.336	0.657
Marginal	0.475	0.512	1.000

LCA Local Independence

- LCA has the property of local independence - that given class, item responses are independent.
- To give an example, consider Items 1 and 2:

	$P(X_j = 1 \mid c)$			
Item j	$c = 1$	$c = 2$	$c = 3$	$c = 4$
1	0.512	0.915	0.306	0.758
2	0.657	0.979	0.307	0.872

	Item 1		
Item 2	0	1	Marginal
0	0.002	0.019	0.021
1	0.083	0.896	0.979
Marginal	0.085	0.915	1.000

LCA Local Independence

- LCA has the property of local independence - that given class, item responses are independent.
- To give an example, consider Items 1 and 2:

	$P(X_j = 1 \vert c)$			
Item j	$c = 1$	$c = 2$	$c = 3$	$c = 4$
1	0.512	0.915	0.306	0.758
2	0.657	0.979	0.307	0.872

	Item 1		
Item 2	0	1	Marginal
0	0.481	0.212	0.693
1	0.213	0.094	0.307
Marginal	0.694	0.306	1.000

LCA Local Independence

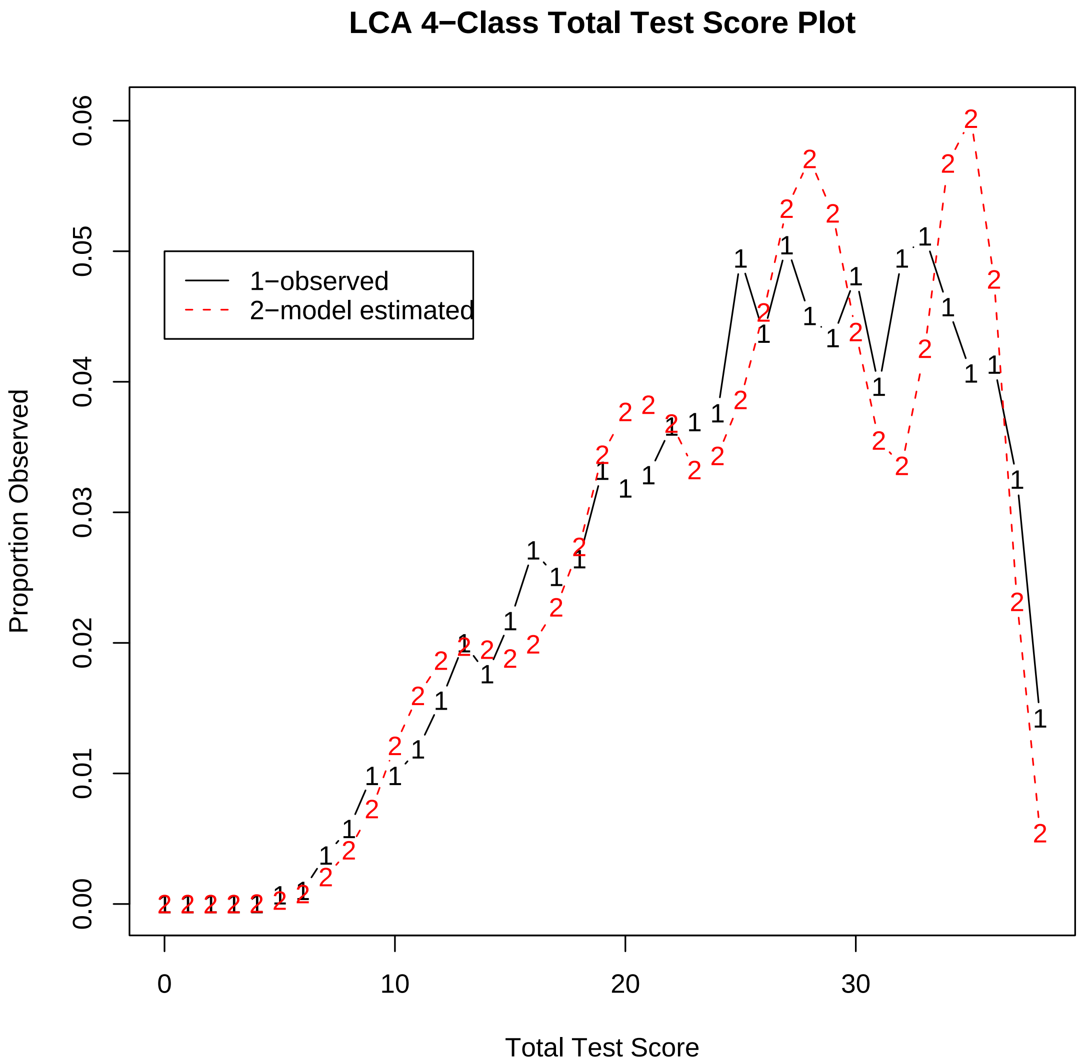
- LCA has the property of local independence - that given class, item responses are independent.
- To give an example, consider Items 1 and 2:

	$P(X_j = 1 \mid c)$			
Item j	$c = 1$	$c = 2$	$c = 3$	$c = 4$
1	0.512	0.915	0.306	0.758
2	0.657	0.979	0.307	0.872

	Item 1		
Item 2	0	1	Marginal
0	0.031	0.097	0.128
1	0.211	0.661	0.872
Marginal	0.242	0.758	1.000

Validation - Total Test Score

As a measure of validation, consider examining the total test score as observed and predicted by our model estimates:



LCA Limitations

- LCA has limitations which make its application to educational measurement difficult:
  - Classes not known prior to analysis.
  - Class characteristics not know until after analysis.
- Both of these problems are related to LCA being an exploratory procedure for understanding data.
- CDMs can be thought of as confirmatory versions of LCA.
  - By placing constraints on the class item probabilities and specifying what our classes mean prior to analysis.

Notation

I denotes the total number of examinees.

J denotes the total number of items.

K denotes the total number of attributes.

Q has elements q_{jk} that indicate whether mastery of the k^{th} attribute is required by the j^{th} item

$q_{j\cdot}$ denotes the number of attributes measured by item j:

$$q_{j\cdot} = \sum_{k=1}^{K} q_{jk}$$

$\boldsymbol{\alpha}_i$ is a set of indicators of attribute mastery for examinee i for all K attributes.

CDM are a Subset of Latent Class Models

A latent class model for the response vector of individual i:

$$P(\mathbf{X}_i = \mathbf{x}_i) = \sum_{\boldsymbol{\alpha}} \overbrace{P(\boldsymbol{\alpha})}^{\text{Structural}} \prod_{j=1}^{J} \overbrace{P(X_{ij} = x_{ij}|\boldsymbol{\alpha})}^{\text{Measurement}}$$

- For K attributes, total of 2^K classes are defined.
- Equality constraints determined by:
  - Choice of model.
  - Q-matrix specifications.
- Basic CDMs assume local independence conditional on skill pattern.

Common CDMs

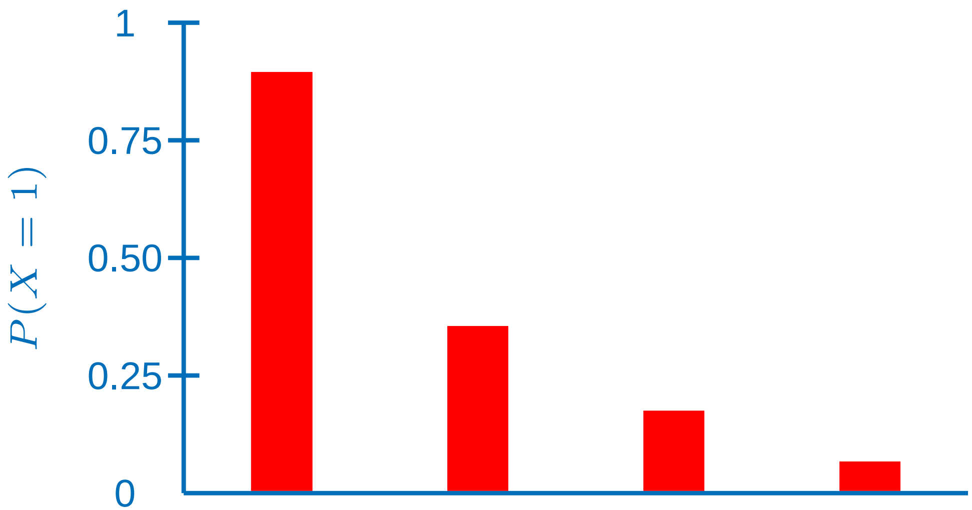
- The DINA model has two classes per item (mastered everything or not).
- The NIDA model constrains the penalty for lacking an attribute to be equal across all items.
- To parsimoniously parameterize the Unified Model (DiBello, Stout, & Roussos, 1995), the Reparameterized Unified Model (or RUM; Hartz, 2002) was developed.
- The RUM blends the two features of the DINA and NIDA models:
  - Parameterization of item-specific parameters (like the DINA).
  - Parameterization of multiple response classes per item (like the NIDA).

RUM Demonstration

- Consider the item:

$$2 + 3 - 1$$

- Addition and Subtraction are skills needed to correctly answer this problem.



Add	✓	✓		
Sub	✓		✓	

Reparameterized Unified Model (Hartz, 2002)

$$P\left(X_{ij}=1|\boldsymbol{\alpha}_i\right)=\pi_j^*\prod_{k=1}^{K} r_{jk}^{(1-\alpha_{ik})\times q_{jk}}$$

- Shown is the reduced form of the RUM (or REDRUM?), which is commonly used in practical applications.
- $\mathbf{q}_j$ is the pre-specified row vector $(1\times K)$ of Q-matrix entries for item j.
  - The RUM places $2^{q_{j\cdot}}$ equality constraints on the 2^K class item response probabilities.
- π_j^* is the maximum probability of correct response conditional on mastery of all Q-matrix attributes for item j.
- r_{jk}^* is the “penalty” imposed for missing attribute k.

Reduced Rum Item Response Constraints

Total Attributes: 4

Total Classes: 16

Model: Reduced RUM

Q-matrix:

	α_1	α_2	α_3	α_4
Item	Add	Sub	Mult	Div
$1.)\ 2+3-1$	1	1	0	0
$2.)\ 4/2$	0	0	0	1
$3.)\ 3 \times (4-2)$	0	1	1	0

Equivalence Classes Per Item: $(2^{q_{j\cdot}})$

Item	1	2	3
Classes	4	2	4

Equal $P(X_{ij} = 1|\boldsymbol{\alpha}_i)$ Denoted By Color

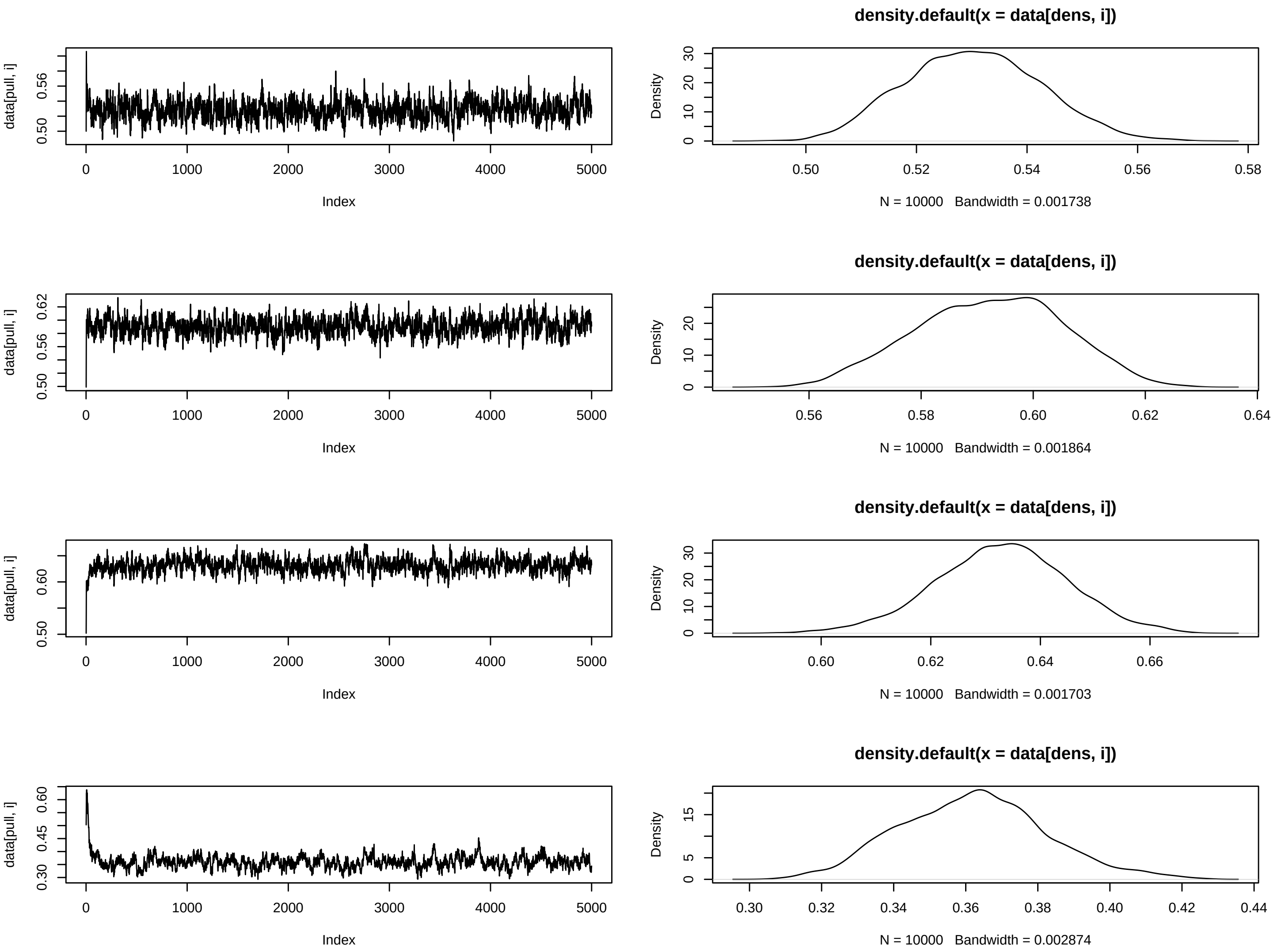
$\boldsymbol{\alpha}$	Item 1	Item 2	Item 3
[0000]	$\pi_1^* r_{11}^* r_{12}^*$	$\pi_2^* r_{24}^*$	$\pi_3^* r_{32}^* r_{33}^*$
[0001]	$\pi_1^* r_{11}^* r_{12}^*$	π_2^*	$\pi_3^* r_{32}^* r_{33}^*$
[0010]	$\pi_1^* r_{11}^* r_{12}^*$	$\pi_2^* r_{24}^*$	$\pi_3^* r_{32}^*$
[0011]	$\pi_1^* r_{11}^* r_{12}^*$	π_2^*	$\pi_3^* r_{32}^*$
[0100]	$\pi_1^* r_1^*$	$\pi_2^* r_{24}^*$	$\pi_3^* r_{33}^*$
[0101]	$\pi_1^* r_1^*$	π_2^*	$\pi_3^* r_{33}^*$
[0110]	$\pi_1^* r_1^*$	$\pi_2^* r_{24}^*$	π_3^*
[0111]	$\pi_1^* r_1^*$	π_2^*	π_3^*
[1000]	$\pi_1^* r_2^*$	$\pi_2^* r_{24}^*$	$\pi_3^* r_{32}^* r_{33}^*$
[1001]	$\pi_1^* r_2^*$	π_2^*	$\pi_3^* r_{32}^* r_{33}^*$
[1010]	$\pi_1^* r_2^*$	$\pi_2^* r_{24}^*$	$\pi_3^* r_{32}^*$
[1011]	$\pi_1^* r_2^*$	π_2^*	$\pi_3^* r_{33}^*$
[1100]	π_1^*	$\pi_2^* r_{24}^*$	$\pi_3^* r_{33}^*$
[1101]	π_1^*	π_2^*	$\pi_3^* r_{33}^*$
[1110]	π_1^*	$\pi_2^* r_{24}^*$	π_3^*
[1111]	π_1^*	π_2^*	π_3^*

RUM Estimation

- To estimate the parameters of the RUM, an MCMC algorithm was created.
- *Arpeggio* was created by Hartz (2002).
- Subsequent modifications by Henson and Templin (2003 - 2005) improved the efficiency and consistency of the algorithm and parameter estimates.
- With an MCMC algorithm the first thing to do following an analysis is to check convergence.
- To the user, this might seem as though we are validating the result.
- As we will see, model convergence does not correspond to a valid result.

MCMC Convergence Check

The timeseries of the MCMC chain will indicate convergence:



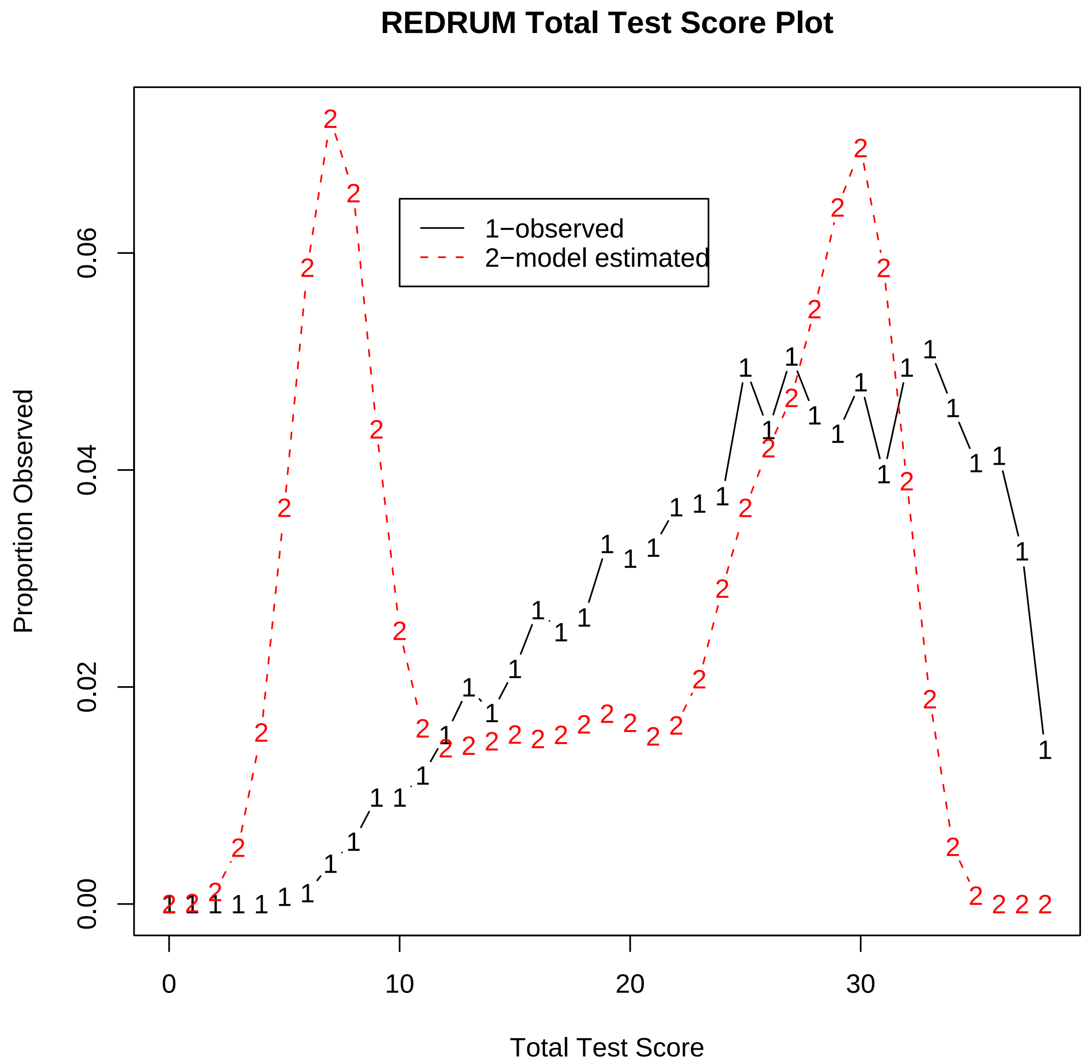
REDRUM Item Parameter Estimates

Reading Data: REDRUM Item Parameters

Item	π^*	r_1^*	r_2^*	r_3^*	r_4^*
1	0.857	0.543	0	0	0
2	0.926	0	0	0.539	0
3	0.947	0.628	0	0	0
4	0.668	0	0.509	0	0
5	0.958	0	0	0.688	0.845

Validation - Total Test Score

As a measure of validation, consider examining the total test score as observed and predicted by our model estimates:



Item Completeness

- The quality of cognitive diagnosis model estimates are partially determined by the accuracy of the entries in the Q-matrix.
- In practice, situations occur where the Q-matrix is not accurately specified.
- For each item, the RUM maps the misspecified Q-matrix entries onto a single latent continuum.
- The RUM incorporates a continuous examinee parameter, denoted by θ_i, that represents the examinee's latent "ability" across the misspecified attributes.

Reparameterized Unified Model (Hartz, 2002)

$$P\left(X_{ij}=1|\boldsymbol{\alpha}_i,\theta_i\right)=P_{c_j}(\theta_i)\left(\pi_j^*\prod_{k=1}^{K}r_{jk}^{(1-\alpha_{ik})\times q_{jk}}\right)$$

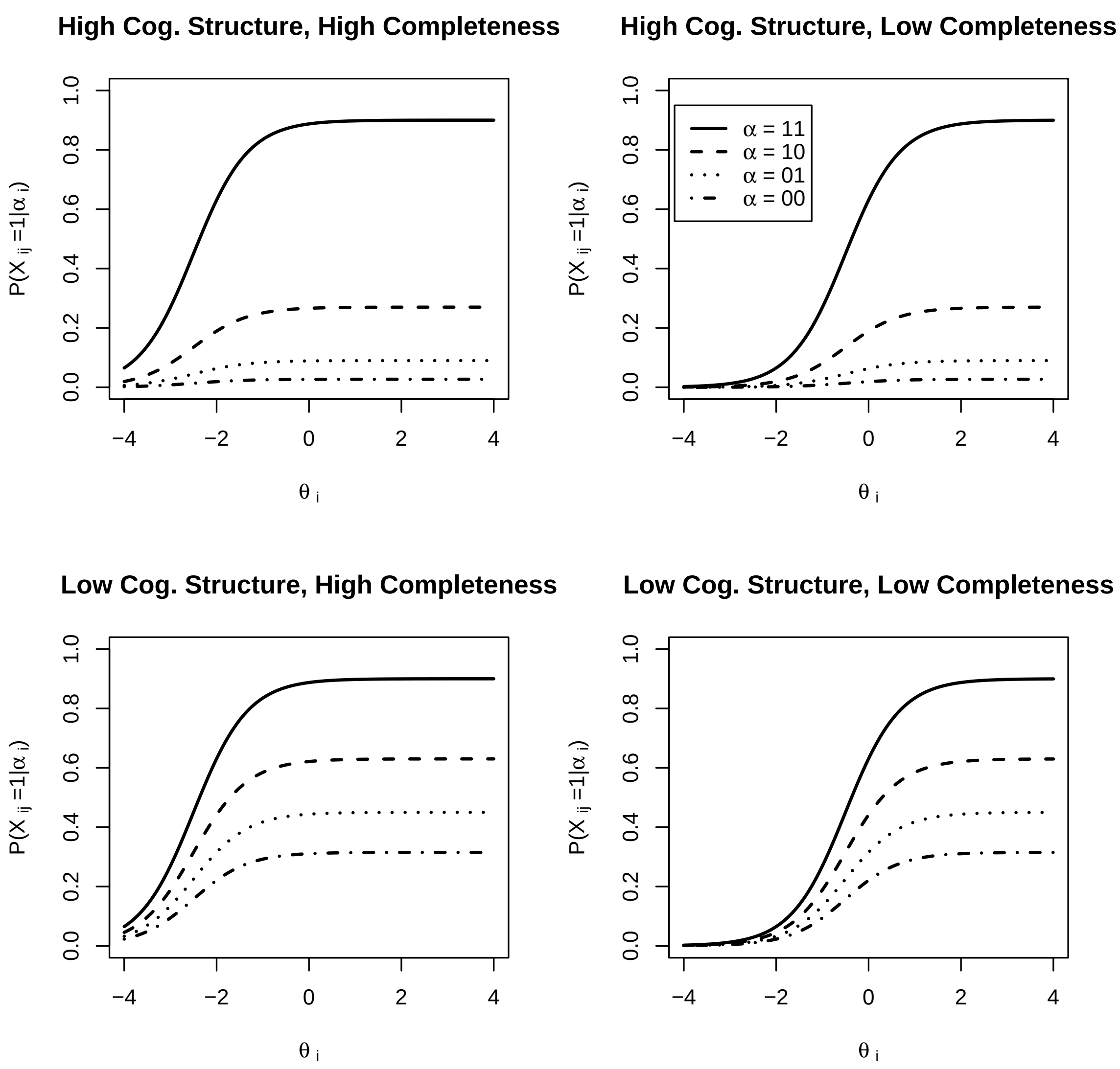
- The π^* and r^* terms function as in REDRUM.
- The "completeness" term:

$$P_{c_j}(\theta_i)=\left(\frac{e^{1.701(c_j+\theta_i)}}{1+e^{1.701(c_j+\theta_i)}}\right)$$

- The c parameter is similar to the difficulty parameter in a Rasch model ($c=-b$).
- In practice, the completeness parameter is difficult (if not impossible) to estimate.
- Incompleteness only penalizes the examinee.

RUM Illustrated

Consider an item requiring two attributes $(q_{j\cdot} = 2)$:



More on Completeness

- The completeness term relaxes the LI assumption of the LCA portion of the model.
- Inclusion of the completeness term makes the RUM more like a Finite Mixture Model.
- Because of the Rasch model parameterization, conditional on class, the RUM fixes the tetrachoric correlation between items to be approximately 0.5.
  - The tetrachoric correlation between items given class is the same for all pairs of items and for all classes.
  - Perhaps a cause in the problems estimating the model?

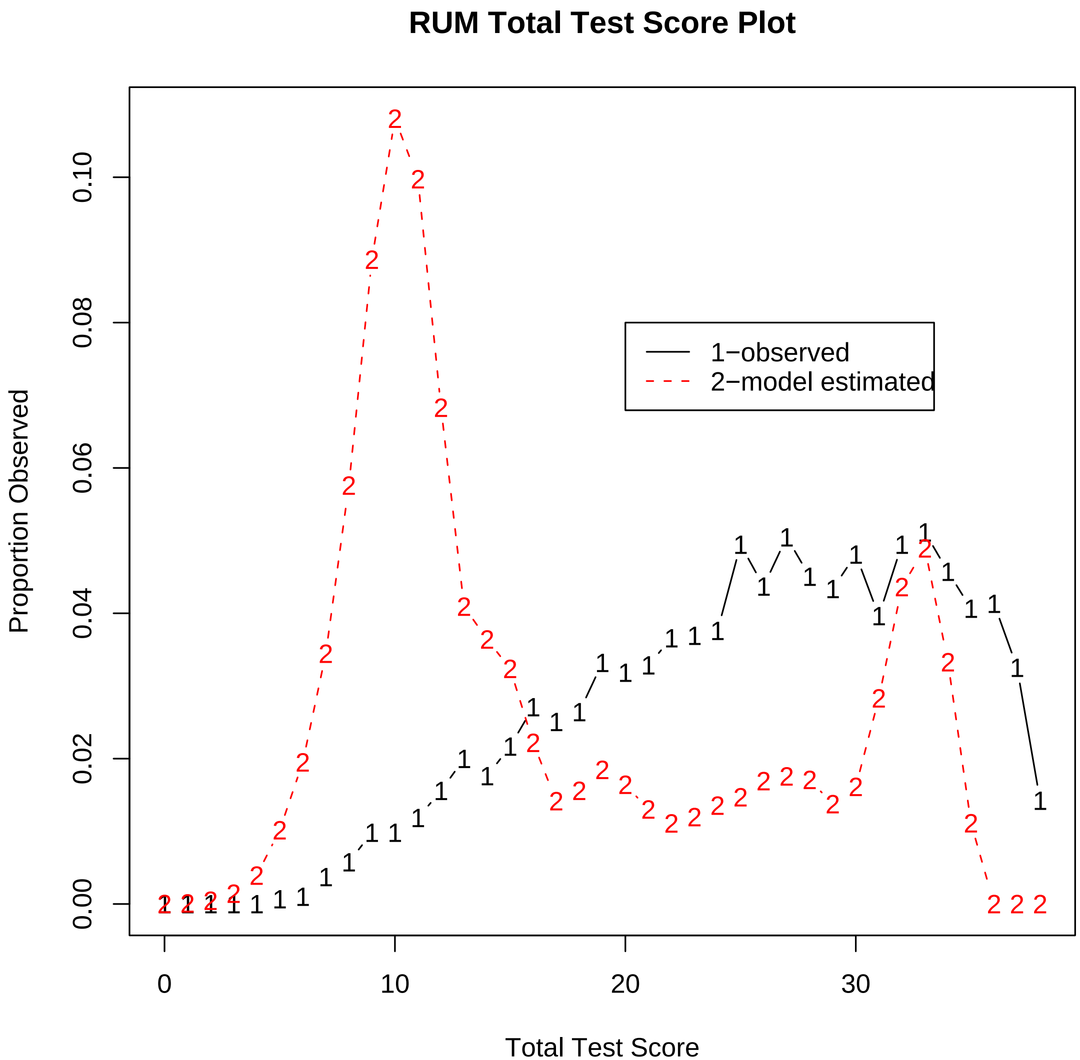
RUM Item Parameter Estimates

Reading Data: RUM Item Parameters

Item	π^*	r_1^*	r_2^*	r_3^*	r_4^*	c
1	0.932	0.848	0	0	0	1.213
2	0.993	0	0	0.959	0	1.198
3	0.991	0.978	0	0	0	1.227
4	0.799	0	0.656	0	0	1.529
5	0.983	0	0	0.956	0.909	1.505

Validation - Total Test Score

As a measure of validation, consider examining the total test score as observed and predicted by our model estimates:



The Search for a Better Model

- Because of the difficulties involved in parameterizing and estimating the RUM, a new model was needed.
- The new model would have to be one that would be sensitive to item incompleteness.
- It would also be nice if the new model did not add many more parameters.
- The result of the process was the development of the Random Effects Reparameterized Unified Model (or RERUM; Templin & Henson, 2005).

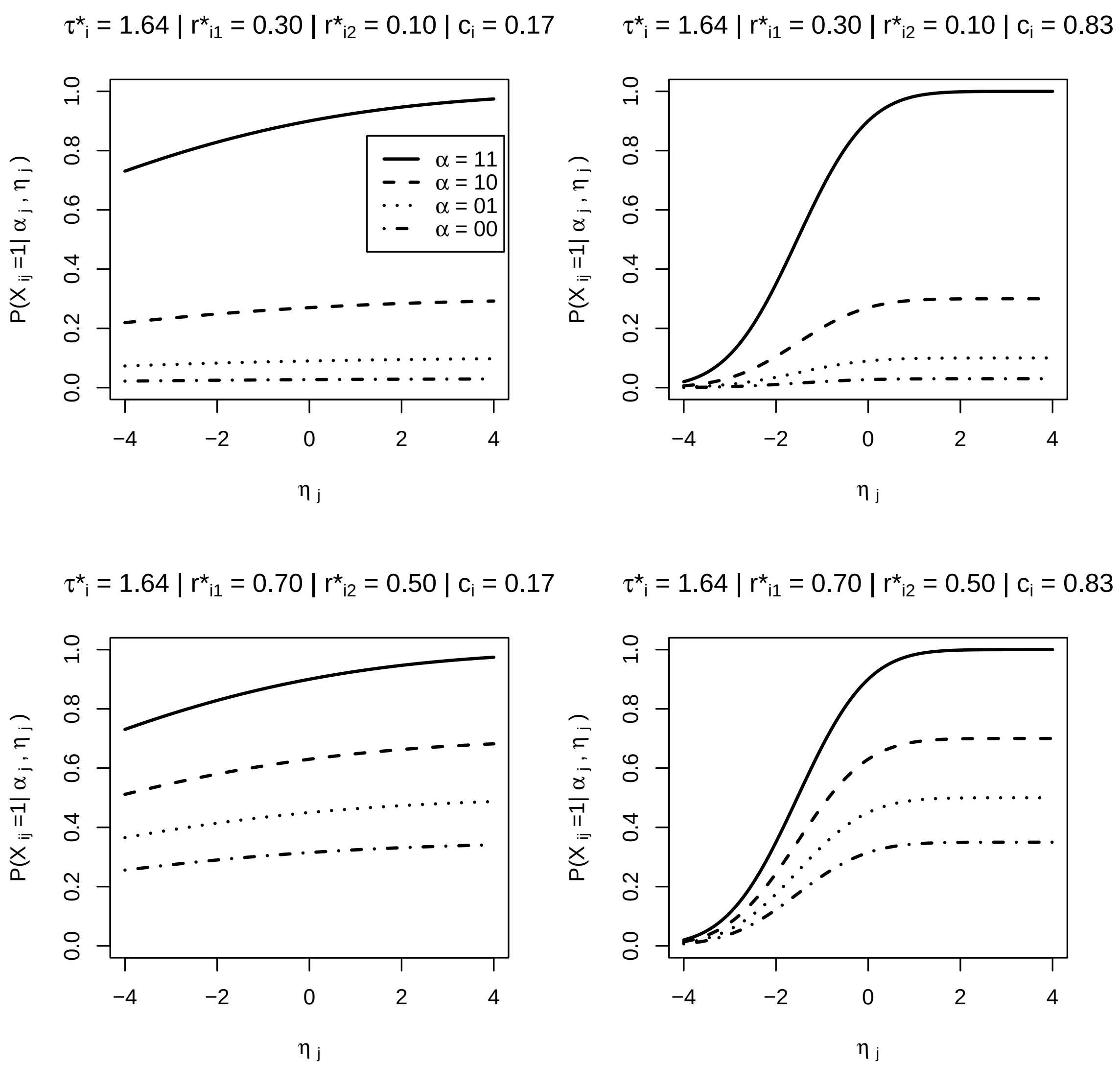
Random Effects RUM

$$P\left(X_{ij}=1|\boldsymbol{\alpha}_i,\theta_i\right)=\Phi\left(\tau_j+c_j\theta_i\right)\prod_{k=1}^{K} r_{jk}^{(1-\alpha_{ik})\times q_{jk}}$$

- The r^* functions as in RUM.
- The $\Phi(\cdot)$ function is the standard normal CDF.
- c ranges from zero (item is complete) to one (item is incomplete).
  - c could also go to negative one.
- The τ_j is the intercept parameter
- Same number of parameters as the RUM.

RERUM Illustrated

Consider an item requiring two attributes $(q_{j\cdot} = 2)$:



Conditional Correlations

- The RERUM completeness term also relaxes the LI assumption of the LCA portion of the model.

- Conditional on class, the RERUM fixes the tetrachoric correlation between items to be the product of the c parameters:

$$\rho_{x_j,x_k|\boldsymbol{\alpha}} = c_j c_k$$

- The tetrachoric correlation between items conditional on class is now estimated as part of the model.

- Items with high conditional correlations may indicate places to modify Q-matrix.

RERUM Item Parameter Estimates

Reading Data: RERUM Item Parameters

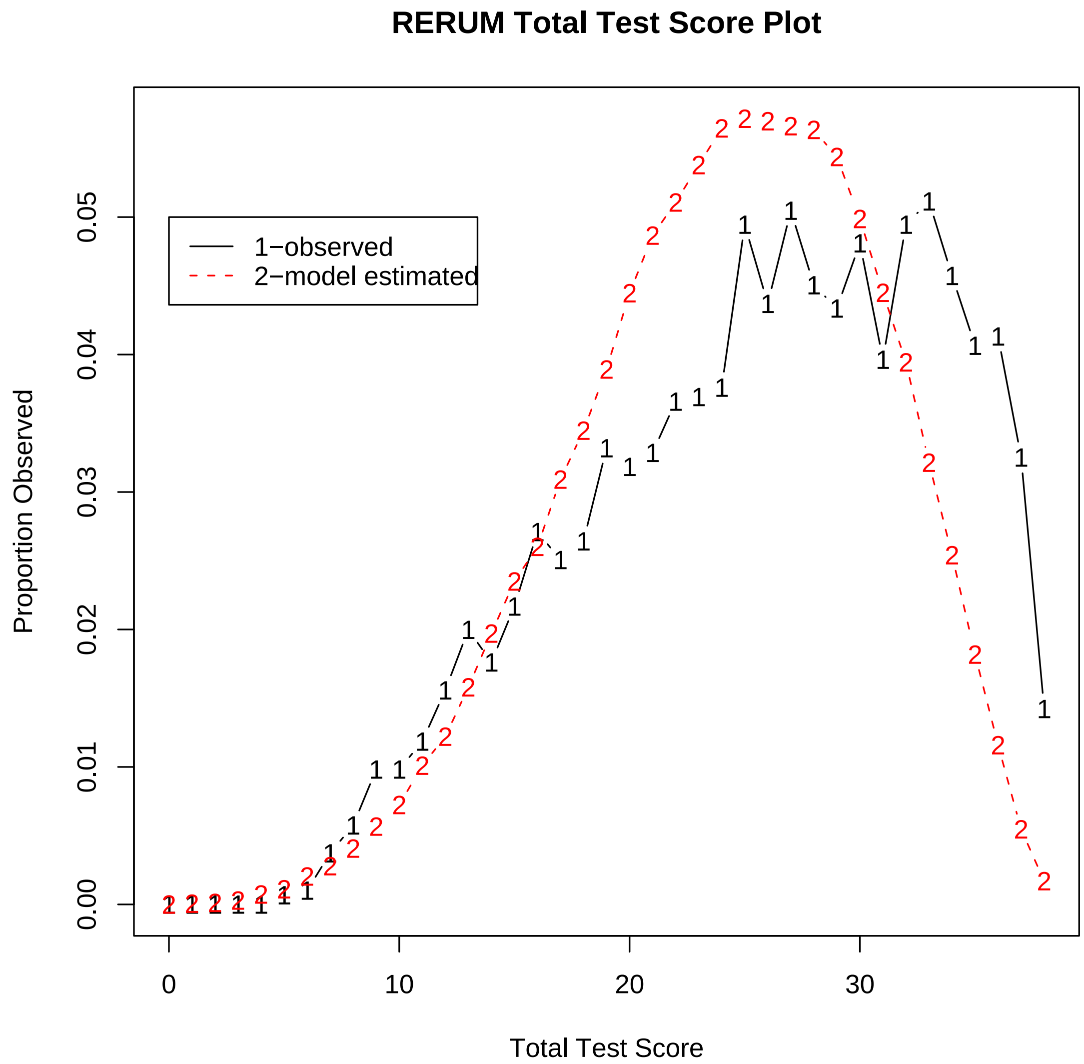
Item	τ	r_1^*	r_2^*	r_3^*	r_4^*	c
1	0.396	0.959	0	0	0	0.495
2	0.668	0	0	0.888	0	0.597
3	0.687	0.966	0	0	0	0.595
4	0.080	0	0.950	0	0	0.405
5	0.684	0	0	0.934	0.890	0.538

Conditional Item Correlations:

	1	2	3	4
2	0.296			
3	0.295	0.355		
4	0.200	0.242	0.241	
5	0.266	0.321	0.320	0.218

Validation - Total Test Score

As a measure of validation, consider examining the total test score as observed and predicted by our model estimates:

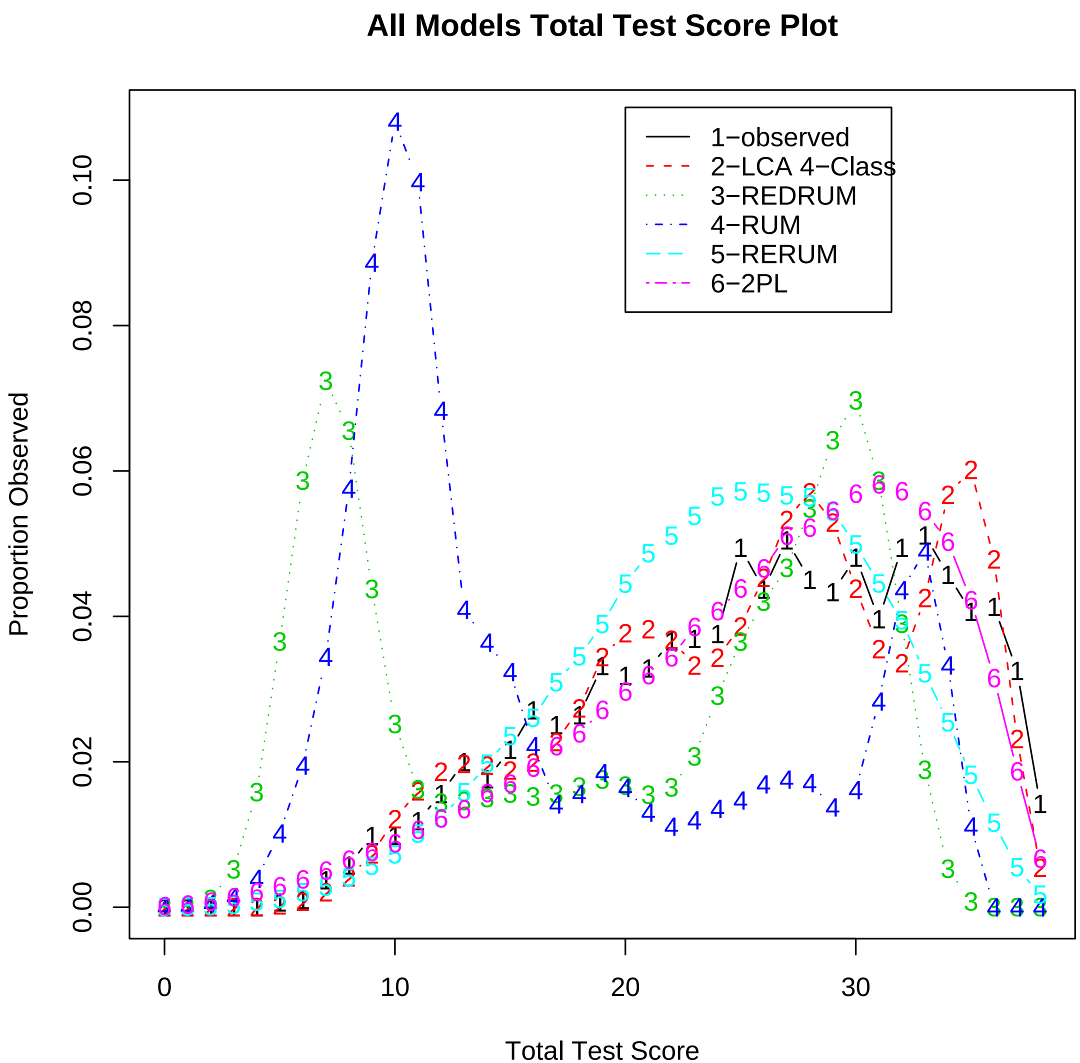


In Summary

- Cognitive diagnosis models have the potential to provide meaningful information to the test-taker, instructor, and legislator.
- However, there are obstacles to overcome when choosing the type of model to be applied.
- Retrofitting of IRT-built tests is not often successful.
- Diagnostic test construction methods do exist (see Henson, 2004).
- Validation should not be forgotten.
- Oh, and about model fit...

Comparing All of Today's Models

Here is a plot of the test score distributions for all of today's models, along with one from a 2PL model:



Additional CDM Topics

- Polytomous item model extensions.
- Polytomous attribute model extensions.
- Different ways of modeling of examinee proficiency space.
- Estimation Methods.
- Model fit.
- Psychological Applications.
- Exploratory methods for Q-matrix discovery.

Concluding Remarks

- Cognitive diagnosis models provide a method for estimating latent skills.
- Such models are in their infancy - time will tell if these models are effective at giving information about examinees.
- A good way to start using such models is from the beginning - design tests with diagnosis/classification in mind.
- Perhaps the first issue solved will be the debate over skills assessment or cognitive diagnosis...
- Thank you.