Diagnostic Measurement: Theory, Methods, Applications, and Software

2016 NCME Training Session

April 8, 2016

Washington, DC

Training Session Schedule

Time	Topic
8:00am-10:00am	Section 1: Conceptual Foundations of Diagnostic Measurement (Theory; Jonathan)
10:00am-10:15am	*Break #1*
10:15am-11:45am	Section 2: Fundamental Concepts of DCMs (Applications; Jonathan)
11:45am-1:15pm	*Lunch*
1:15pm-3:00pm	Section 3: Psychometric Models (Methods; Jonathan)
3:00pm-3:15pm	*Break #2*
3:15pm-4:00pm	Section 4: Estimation of DCMs using Mplus (Software; Meghan)
4:00pm-4:30pm	Section 5: Issues in Software and Estimation (Software; Hongling)
4:30pm-5:00pm	Q & A

About Me…

Jonathan Templin, Ph. D.

Associate Professor

Department of Educational Psychology

University of Kansas

Email:

jtemplin@ku.edu

Website:

http://jonathantemplin.com

Please feel free to contact me with any questions about the material, comments, or suggestions to make things better!

Assisting Me…

Meghan Sullivan

University of Kansas

meg.sullivan@ku.edu

Hongling Lao

University of Kansas

lao@ku.edu

Testing…

- It has been said that the only things certain in life are death and taxes…
  - I propose to add taking tests to that list

- You take a test – you get a score
  - Sometimes how you get your score is obvious
    - The number of items you answered correctly
  - Sometimes it is less obvious
    - A number appears on a screen or in your mailbox after a few weeks
  - Sometimes your score follows you for years

- Often your score gets evaluated by some cut-point
  - You are eligible for admission to grad school
  - You are proficient at mathematics
  - You can begin practicing nursing…or better yet, brain surgery

…Revisited

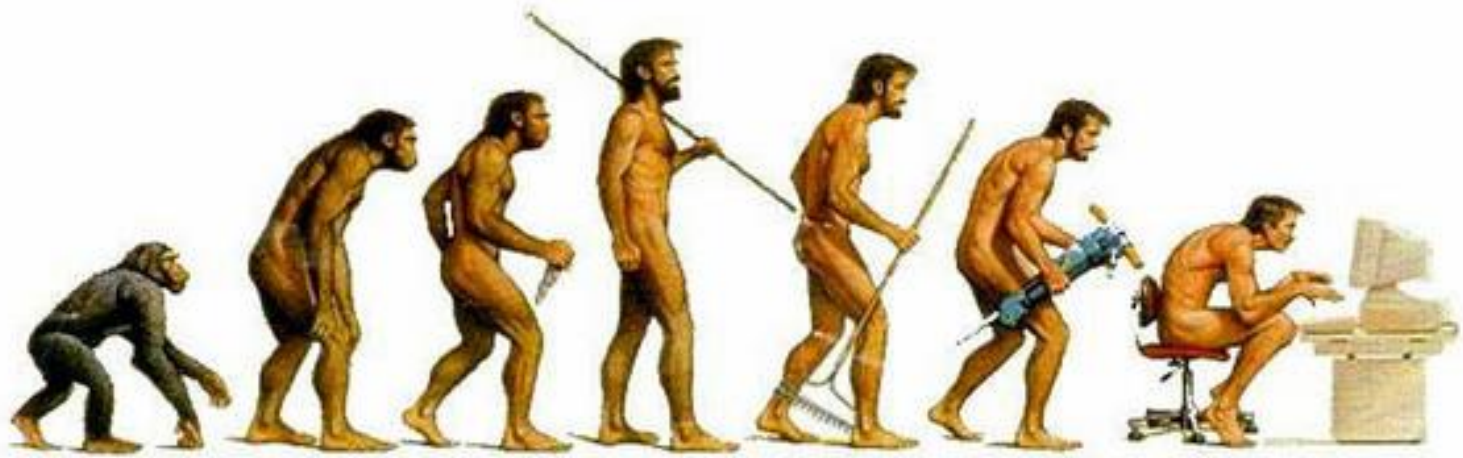
- It is almost unheard of to think about what a test would be like if you didn't get some type of numerical score
  - However, that is the world in which I live

- What if tests were made that didn't just give you a score?
  - Perhaps the test would give you a better sense of direction
    - What to study next
  - Perhaps the test could be shorter
    - I recall being physically sore the day after I took the GRE general and subject tests!
    - Diagnose your standing relative to external criteria (cut-points)

- Further, what if tests could be developed to do all of this *and* to give you a score?
  - Sometimes a score is still needed

- Over the course of the next few hours, I will discuss the world of testing… without the auspices of a score
  - …the world of **<u>diagnostic measurement</u>**

Talk Overview



Lecture Overview

- Key definitions
- Conceptual example
- Example uses of diagnostic models in education
  - Classroom use (formative assessment)
  - Large-scale testing use (summative assessment)
- Why diagnostic models could be used instead of traditional classification methods
- Concluding remarks

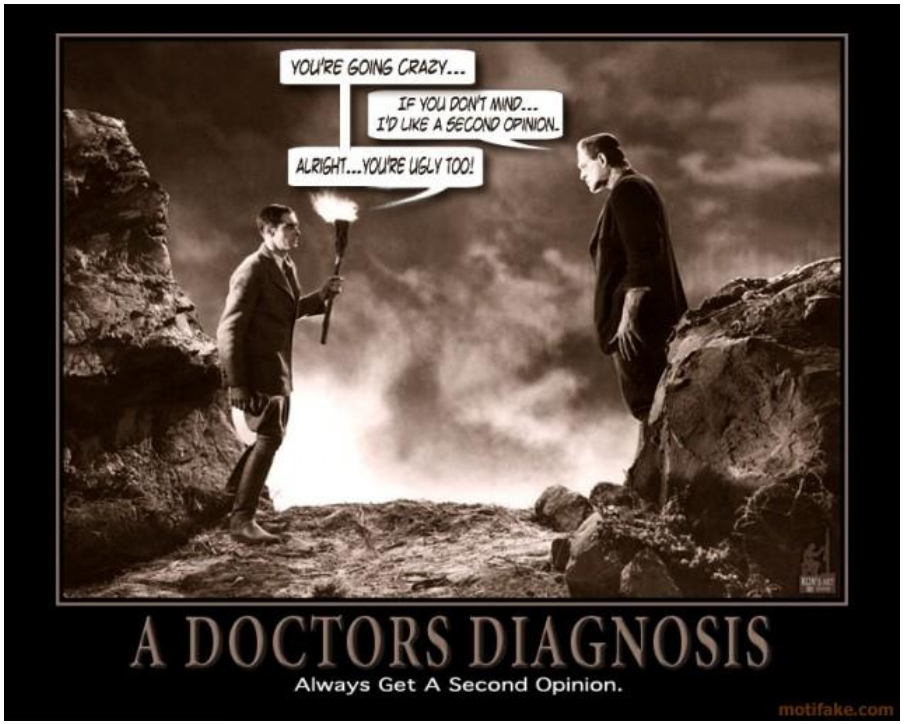
KEY DEFINITIONS

What are Diagnoses?

- The word and meaning of diagnosis is very commonly used in language

- The roots of the word diagnosis:
  - *gnosis*: to know
  - *dia*: from two

- Meaning of diagnoses are deeply ingrained in our society
  - Seldom merits a second thought



Definitions

- *American Heritage Dictionary* definition of *diagnosis*:
  - Generally
    - (a) A critical analysis of the nature of something
    - (b) The conclusion reached by such analysis
  - Medicine
    - (a) The act or process of identifying of determining the nature and cause of a disease or injury through evaluation of a patient's history, examination, and review of laboratory data
    - (b) The opinion derived from such an evaluation
  - Biology
    - (a) A brief description of the distinguishing characteristics of an organism, as for taxonomic classification (p. 500)

Diagnosis: Defined

- A diagnosis is the *decision* that is being made based on information

- Within psychological testing, providing a test score gives the information that is used for a diagnosis

  - BUT, the score is not the diagnosis

  - For this workshop, a diagnosis is by its nature *discrete*
    - Classification

Day-to-Day Diagnosis

- Decisions happen every day:
  - Decide to wear a coat or bring an umbrella
  - Decide to study
  - Decide what to watch on TV tonight

- In all cases:
  - Information (or data) is collected
  - Inferences are made from data based on what likely true state of reality

Diagnosis (Formalized)

- In diagnostic measurement, the procedures of diagnosis are formalized:
  - We make a set of observations – Usually through a set of test questions
  - Based on these questions we make a decision as to the underlying state (or states) of a person
    - The decision is the diagnosis

Diagnosis (Formalized)

- Diagnoses featured in this lecture:
  - Educational Measurement
    - The competencies (skills) that a person has or has not mastered
      - Leads to possible tailored instruction and remediation

- Also possible (and very useful)
  - Psychiatric Assessment
    - The DSM criteria that a person meets
      - Leads to a broader diagnosis of a disorder
  - Personnel selection
    - Matching traits of individuals to needs of jobs

DCM Book: Terminology

- **Respondents**: The people from whom behavioral data are collected
  - Behavioral data considered test item responses for workshop
  - Not limited to only item responses

- **Items**: Test items used to classify/diagnose respondents

- **Diagnostic Assessment**: The method used to elicit behavioral data

- **Attributes**: Unobserved dichotomous characteristics underlying the behaviors (i.e., diagnostic status)
  - Latent variables linked to behaviors diagnostic classification models

- **Psychometric Models**: Models used to analyze item response data
  - Diagnostic Classification Models (DCMs) is the name of the models used to obtain classifications/diagnoses

Diagnostic Classification Model Names

- Diagnostic classification models (DCMs) have been called many different things
  - Skills assessment models
  - Cognitive diagnosis models
  - Cognitive psychometric models
  - Latent response models
  - Restricted (constrained) latent class models
  - Multiple classification models
  - Structured located latent class models
  - Structured item response theory

Psychometric Soapbox

- DCMs are but a small set of tools that must be adapted for a common purpose
  - Part of a methodological toolbox that is used to classify respondents
  - Should also include content experts and end-users of the diagnoses

- DCMs link empirical observations and respondents characteristics
  - The models are only as good as underlying theories

Motivation for DCMs

- Testing more today that we ever have
  - Accountability movement
- What are we getting out of testing?
  - Often, a single score
  - How useful is this score:
    - To make decisions about students?
    - To reflect students' knowledge base or deficiencies?
    - To inform instruction?
- What if a test didn't give a single score?
  - Instead made decisions about students
    - With respect to multiple, discrete facets of a content area
- A diagnostic classification model is a tool that can be used to make these kinds of decisions

CONCEPTUAL EXAMPLE

Diagnostic Modeling Concepts

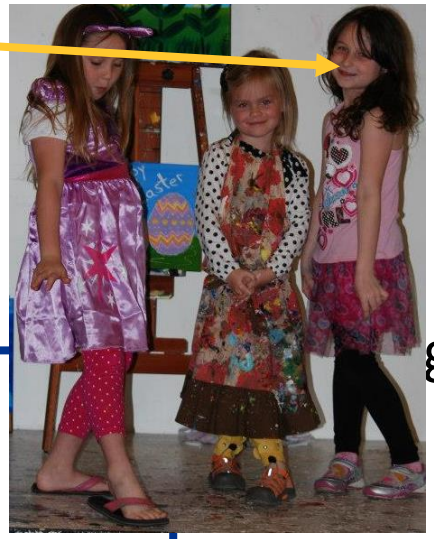
- Imagine that an elementary teacher wants to test basic math ability

- Using traditional psychometric approaches, the teacher could estimate an ability or test score for each respondent
  - Classical Test Theory: Assign respondents a test score
  - Item Response Theory: Assign respondents a latent (scaled) score

- By knowing each respondent‘s score, the students are ordered along a continuum

A Depiction of Traditional Psychometric Estimates

Ordering provided by Daphne Templin



Mathematics Ability

Low

gh

Bieber

Cyrus

Swift

Traditional Psychometrics

- What results is a (weak) ordering of respondents
  - Ordering is called weak because of error in estimates
  - Taylor Swift > Miley Cyrus > Justin Bieber

- Questions that traditional psychometrics cannot answer:
  - Why is The Biebs so low?
    - How can we get him some help?
  - How much ability is “enough” to pass?
    - How much is enough to be proficient?
  - What math skills have the students mastered?

Multiple Dimensions of Ability

- As an alternative, we could have expressed math ability as a set of basic skills:
  - Addition
  - Subtraction
  - Multiplication
  - Division

Ability from a Diagnostic Perspective

	Has Mastered	Has *Not* Mastered
Addition	  	
Subtraction	 	
Multiplication		 
Division		 

Multiple Dimensions of Ability

- The set of skills represent the multiple dimensions of elementary mathematics ability

- Other psychometric approaches have been developed for multiple dimensions
  - Classical Test Theory - Scale Subscores
  - Multidimensional Item Response Theory (MIRT)

- Yet, issues in application have remained:
  - Reliability of estimates is often poor for most practical test lengths
  - Dimensions are often very highly correlated
  - Large samples are needed to calibrate item parameters in MIRT

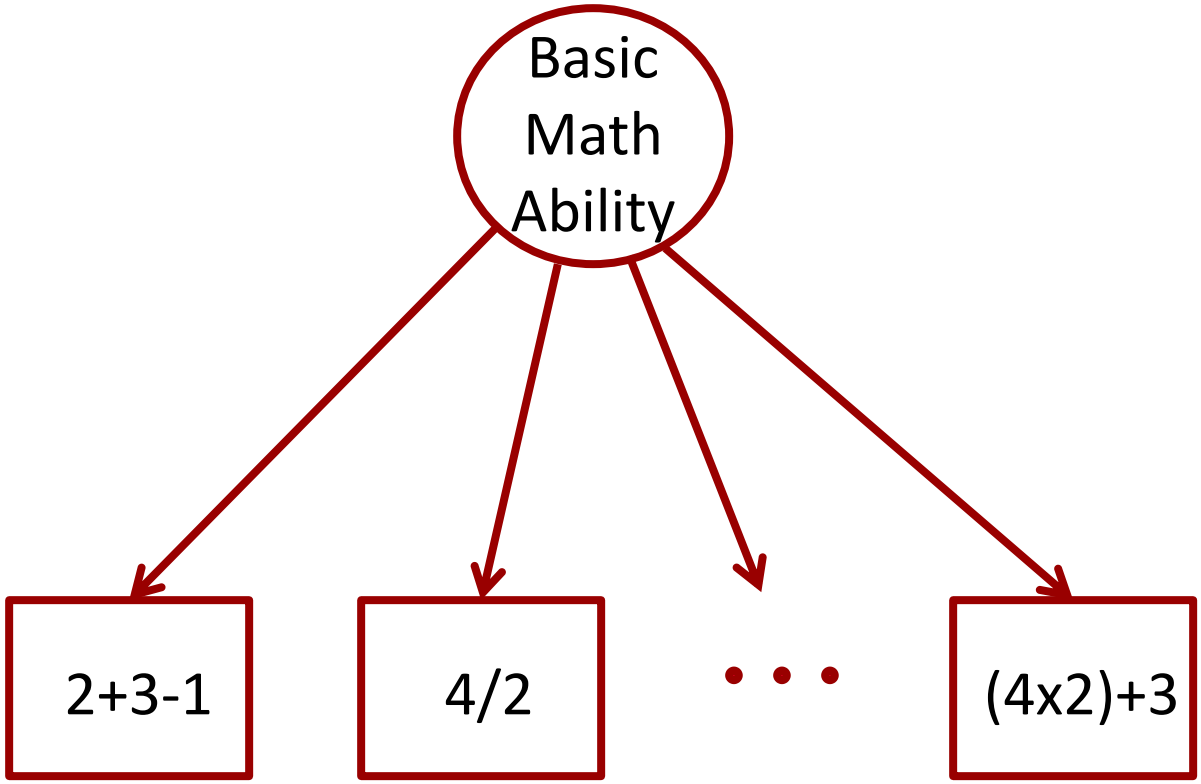
DCMs as an Alternative

- DCMs do not assign a single score

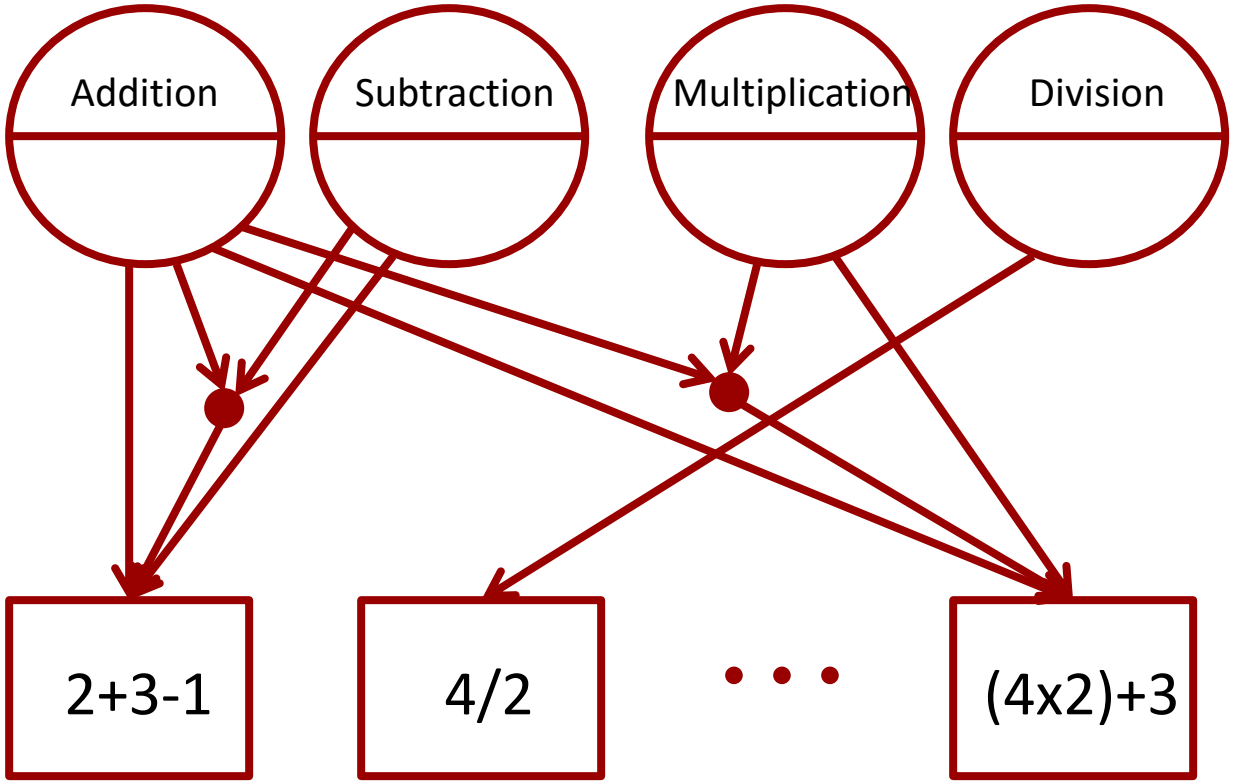
- Instead, a ***profile*** of ***mastered*** attributes is given to respondents
  - Multidimensional models

- DCMs provide respondents valuable information with fewer data demands than other multidimensional models
  - Higher reliability than comparable IRT/MIRT models
  - Complex item structures possible

Path Diagram of Traditional Psychometrics

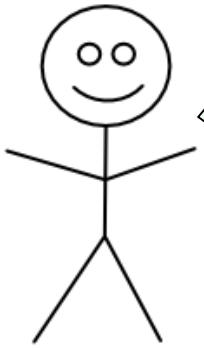


Path Diagram of Diagnostic Models



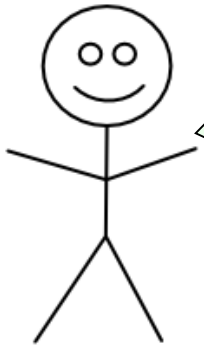
Psychometric Model Comparison

Using Traditional Models



- Has a score of 20
- Has a 75%, a grade of "C"
- Is in the 60th percentile of math
- Scored above the cut off, passes math

Using Diagnostic Models



- *Is proficient* using addition
- *Is proficient* using subtraction
- *Should work* on Multiplication
- *Should work* on Division

DCM Specifics

- Let's expand on the idea of the basic math test
- Possible items may be:
  - 2+3-1
  - 4/2
  - (4 x 2) + 3
- Not all items measure all attributes
- A Q-matrix is used to indicate the attributes measured by each item
  - This is the ***factor pattern matrix*** that assigns the loadings in ***confirmatory factor analysis***

The Q-Matrix

- An example of a Q-matrix using our math test

	Add	Sub	Mult	Div
2+3-1	1	1	0	0
4/2	0	0	0	1
(4 x 2)+3	1	0	1	0

Respondent Profiles

- Respondents are characterized by profiles specifying which attributes have been mastered
  - Numeric values are arbitrary, but for our purposes
    - Mastery given a 1
    - Non-mastery given a 0
- For example:

	Add	Sub	Mult	Div
Respondent A	1	1	0	0

- Respondent profile estimates are in the form of ***probabilities of mastery***

Expected Responses to Items

Q-matrix

	Add	Sub	Mult	Div
2+3-1	1	1	0	0
4/2	0	0	0	1
(4 x 2)+3	1	0	1	0

By knowing which attributes are measured by each item and which attributes have been mastered by each respondent, we can determine the items that will likely be answered correctly by each respondent

Respondent Mastery

	Add	Sub	Mult	Div
Respondent 1	1	1	0	0
Respondent 2	0	1	0	1
Respondent 3	1	0	1	0
Respondent 4	1	1	1	0

Prob Ans #1

Prob Ans #2

Prob Ans #3

Prob Ans #1 & #3

DCM Scoring and Score Reporting

Diagnostic Scoring Report — Student Name: Daphne

Review Your Answers

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Your Answer	✓	✓	✓	✓	a	c	✓	c	d	✓	✓	✓	c	✓	d	a	✓	b	a	a	d	c	b	a	c
Correct Answer	d	a	b	d	d	a	b	d	a	c	a	b	d	c	a	d	a	c	b	d	a	a	a	d	b
Difficulty	e	e	m	m	m	m	h	h	h	m	e	e	m	m	m	h	m	m	h	h	h	h	h	h	h

Score

You correctly answered 10 out of 25 questions.

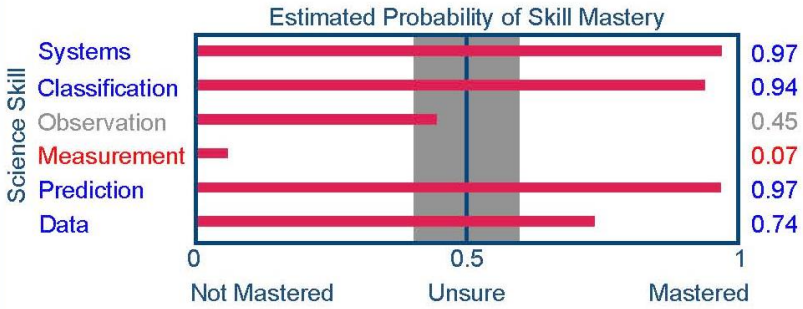
Easy: 4/4; Medium: 5/10; Hard: 1/11

Guide

✓ - Correct answer; o - Omitted answer

e - Easy; m - Medium; h - Hard

Improve Your Skills



Example Questions
3, 14, 2, 17, 19, 23, 9
3, 12, 13, 5, 2, 17, 18, 16, 24, 7
11, 15, 1, 8, 18
22, 20, 10, 11, 5, 6, 18, 25
4, 14, 20, 12, 5, 19, 9
22, 1, 19, 21

from Templin (2007)

DCM Conceptual Summary

- DCMs focus on **WHY** a respondent is not performing well as compared to only focusing on **WHO**

- The models define the chances of a correct response based on the respondent's attribute profile

- Many models have been created ranging in complexity
  - A general DCM: The LCDM
  - The general model subsumes all other latent-variable DCMs

- The model predicts how respondents will answer each item
  - Also allows for classification/diagnoses based on item responses

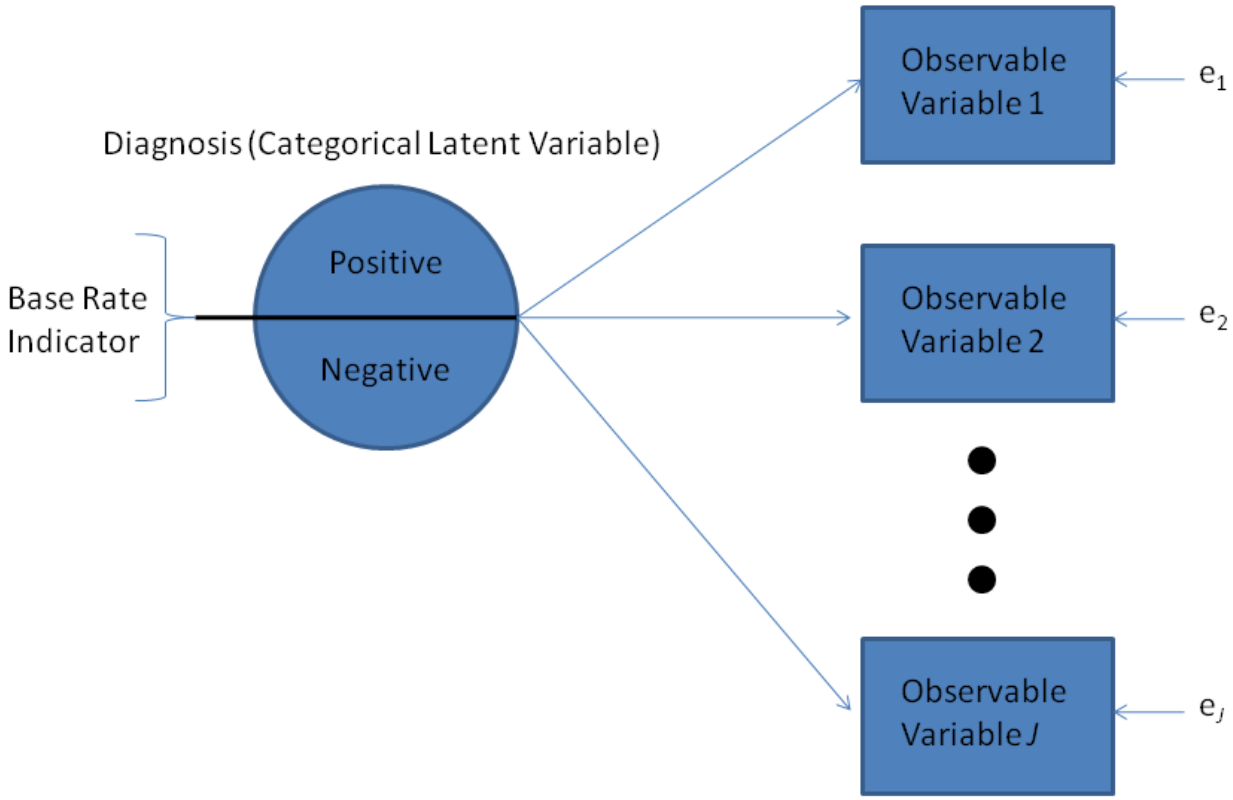
How do DCMs Produce Diagnoses?

- Diagnostic decisions come from comparing observed behaviors to two parts of the psychometric model:

1. Item/variable information (item parameters) — Measurement Model
   - How respondents with different diagnostic profiles perform on a set of test items
   - Helps determine which items are better at discriminating between respondents with differing diagnostic profiles

2. Respondent information pertaining to the base-rate or proportion of respondents with diagnoses in the population — Structural Model
   - Provides frequency of diagnosis (or diagnostic profile)
   - Helps validate the plausibility of the observed diagnostic profiles

Conceptual Model Mapping in DCMs



USES OF DIAGNOSTIC MODEL RESPONDENT ESTIMATES

DCMs In Practice

- To demonstrate the potential benefits of using DCMs, we present a brief example of their use
  - From Henson & Templin (2008); Templin & Henson (2008)

- An urban county in a southern state wanted to improve student's End-Of-Course (EOC) scores on the state's 10^{th} grade Algebra 2 exam

- A benchmark test was given in the middle of a semester
  - Formative test designed to help teachers focus instruction

- Respondents and their teachers received DCM estimates
  - Used these to characterize student proficiency levels with respect to 5 state-specified goals for Algebra 2 (standards)

DCM Study

- The benchmark test was developed for use with a DCM
  - Characteristics of the test were fixed via standard setting
- Five attributes were measured
  - Mastery was defined as meeting the proficient level for each attribute
  - Attributes were largest represented in EOC exam
- Respondents then took the EOC exam
  - 50 item test: Score of 33+ considered proficient
  - Benchmark estimates linked to EOC estimates
- Next slides describe how DCMs can help guide instruction

Descriptive Statistics of Attribute Patterns

- First, the basic descriptive statistics for each possible pattern

- What we expect a respondent with a given attribute pattern to score on the EOC test

Skill Pattern	Expected Score
[00000]	22.9
[00001]	26.0
[00011]	29.3
[00111]	31.4
[01111]	34.8
[11111]	41.9

Gain by Mastery of Each Attribute

- The difference in test score between masters and non-masters of an attribute can be quantified

- Correlation between attribute and EOC score indicates amount of gain in EOC score by mastery of each attribute

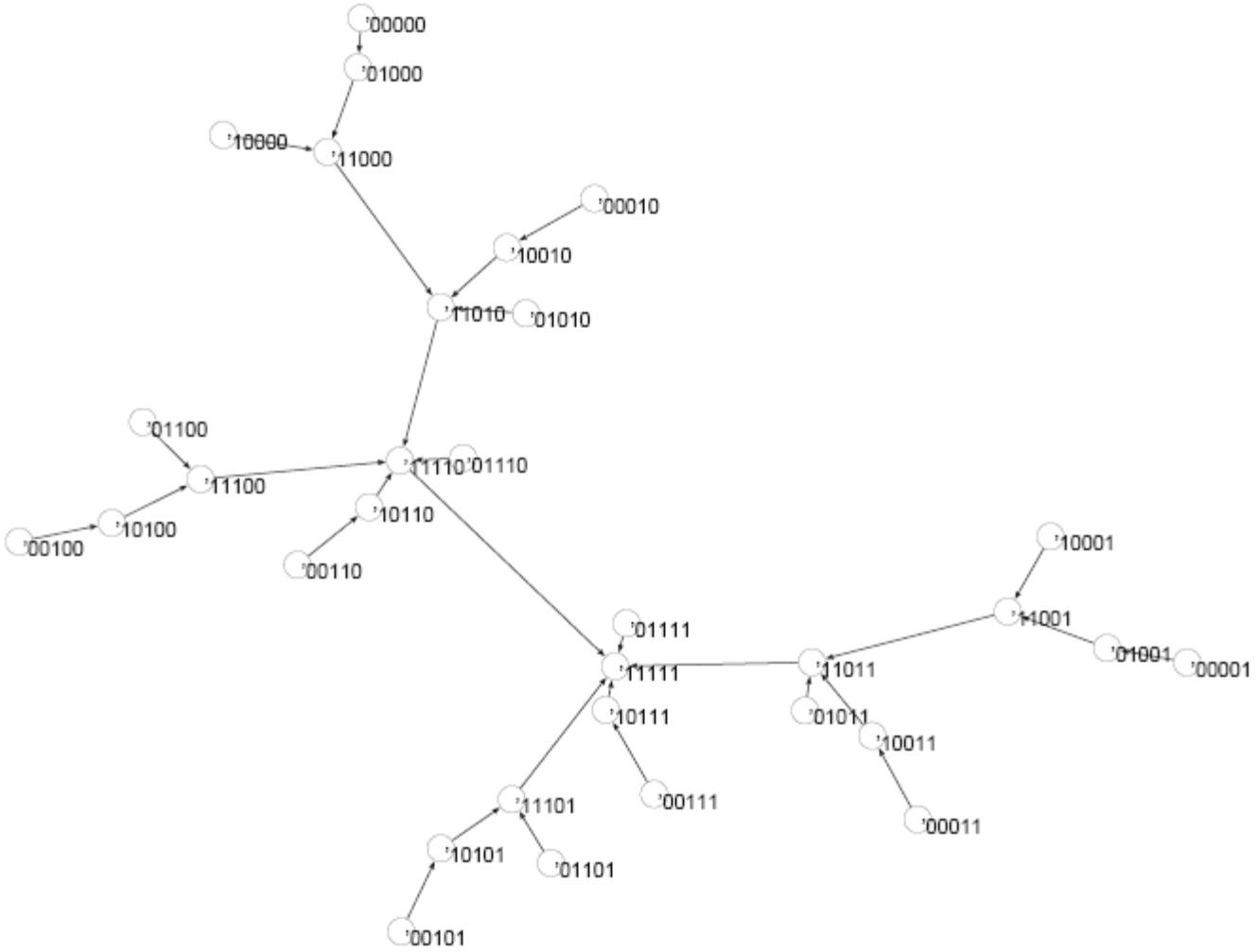
Skill	Gain in Score	Ability Correlation
1	2.61	0.81
2	2.50	0.81
3	1.15	0.63
4	1.19	0.63
5	0.75	0.45

Note: 50 item test

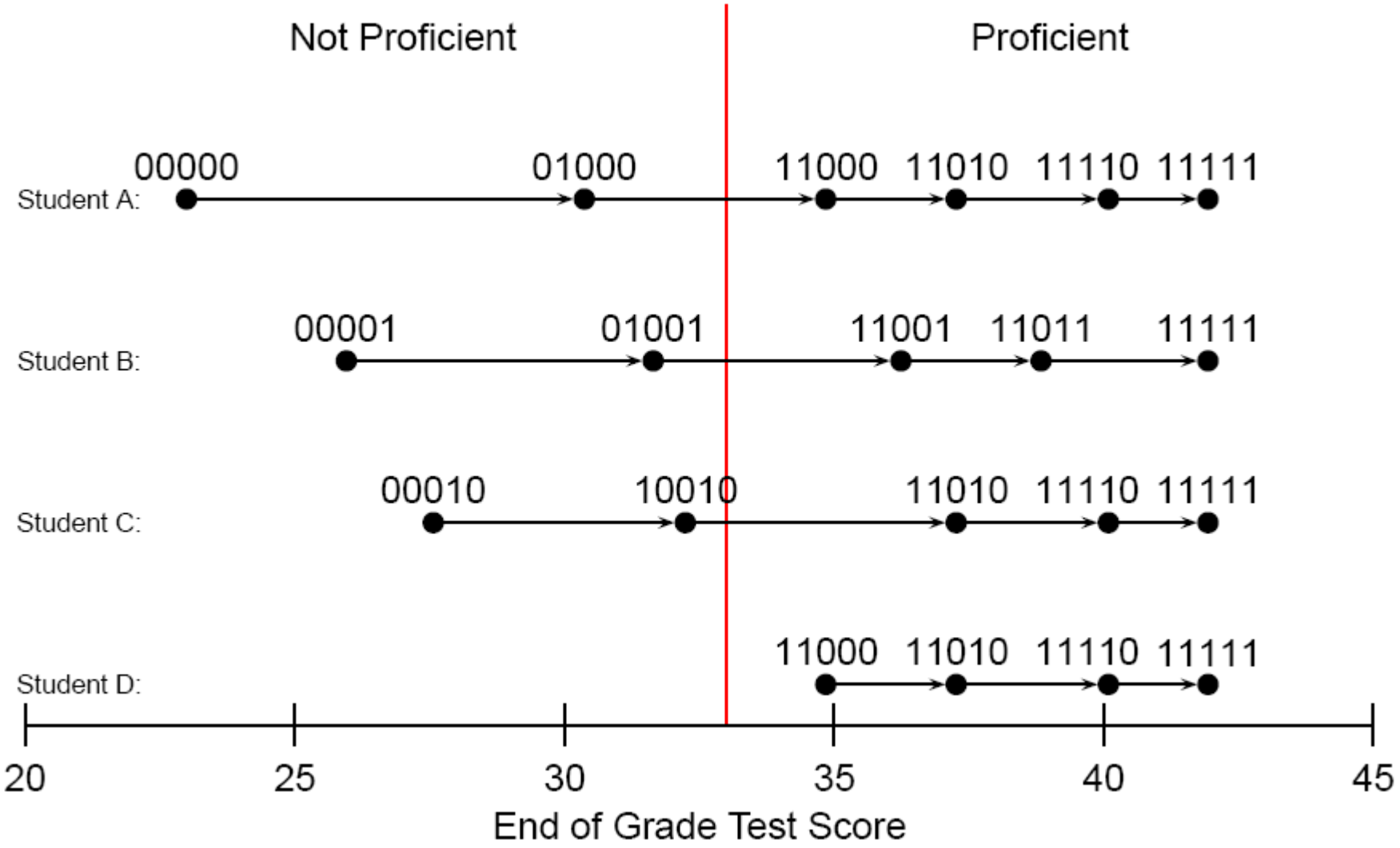
Pathways to Proficiency

- DCMs can be used to form of a "learning path" a respondent can follow that would most quickly lead to proficiency on the EOC test
- The pathway tells the respondent and the teacher the sequence of attributes to learn next that will provide the biggest increase in test score
- This mechanism may help teachers decide focus on when teaching a course
  - Balances time spent on instruction with impact on test score
- Provides a practical implementation of DCMs in today's classroom testing environment

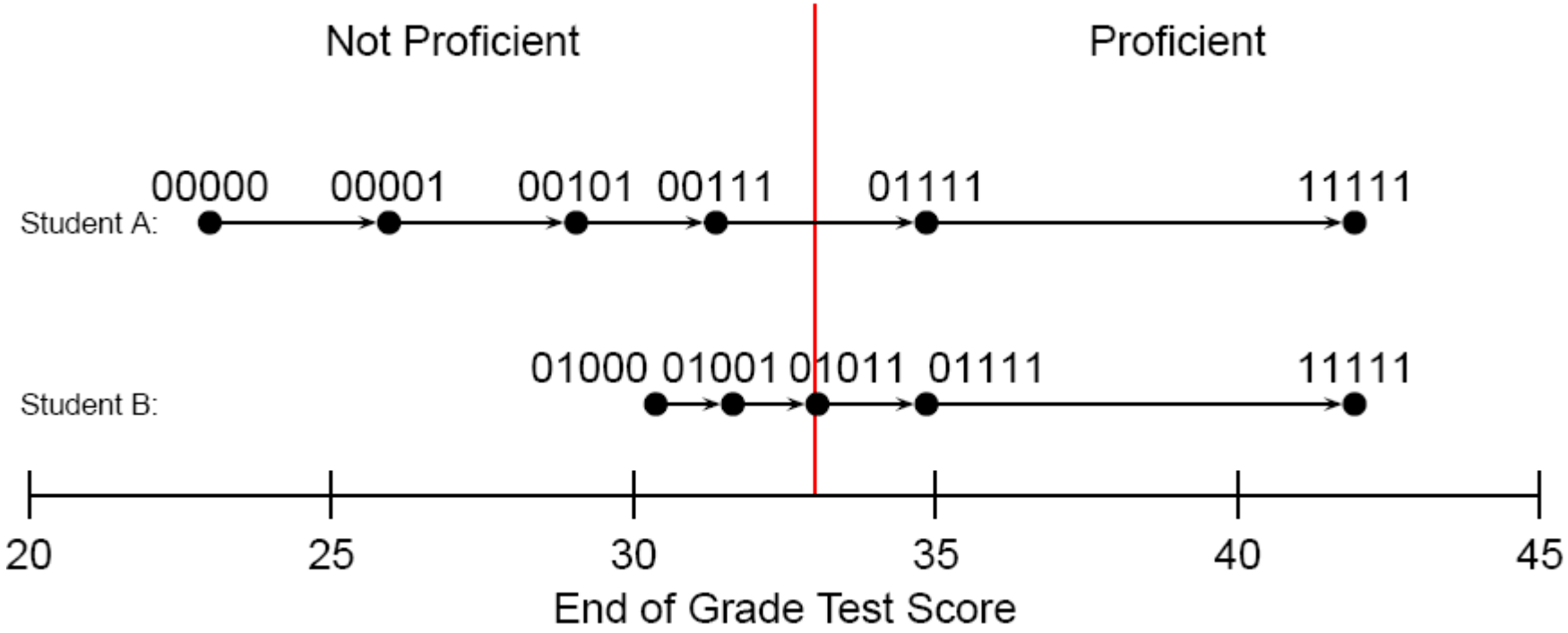
Proficiency Road Map



Fast Path to Proficiency



Harder Paths to Proficiency



- Some paths are less efficient at increasing EOC test scores

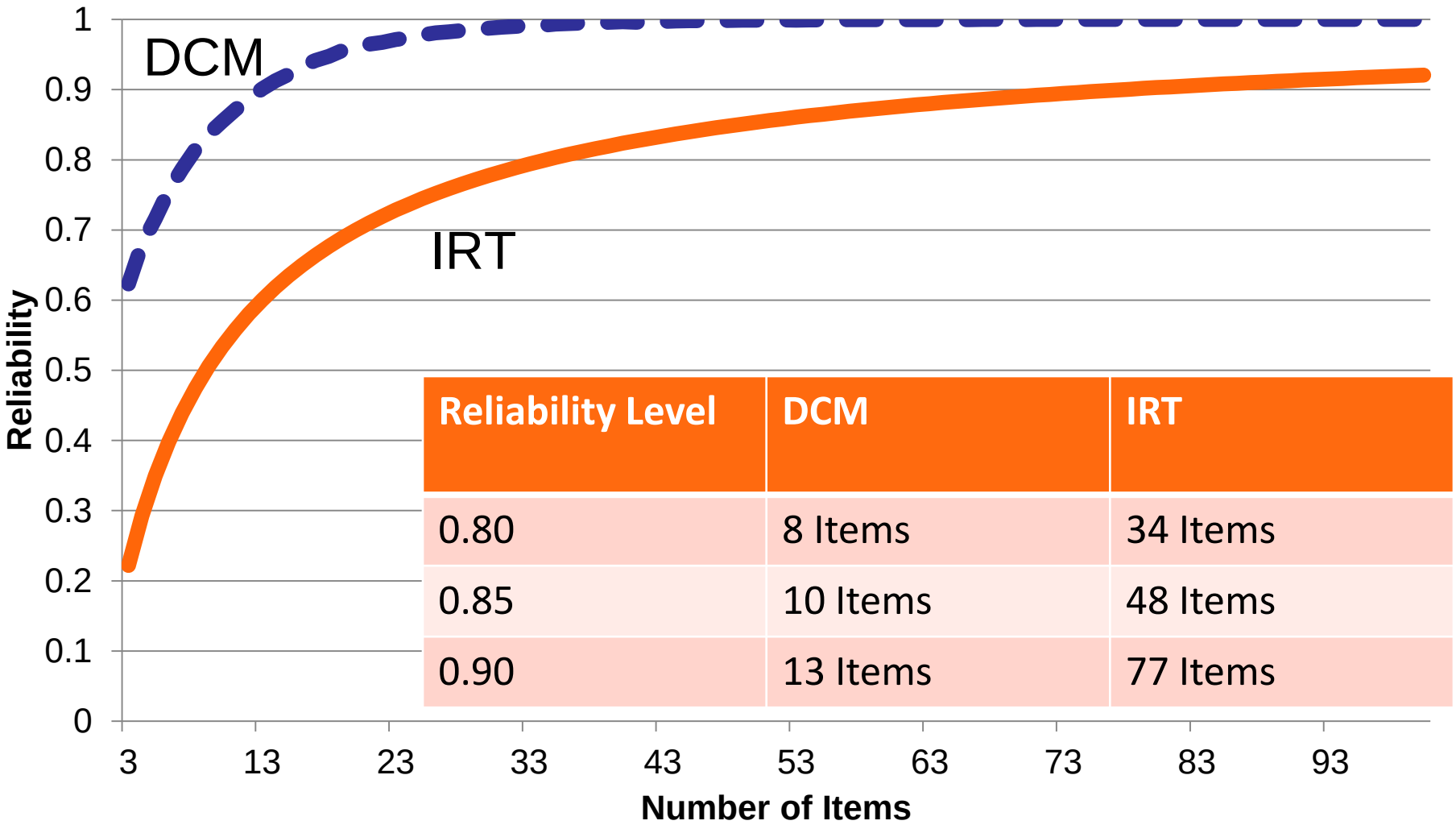
IMPLICATIONS FOR LARGE SCALE TESTING PROGRAMS

DCM Characteristics

- As mentioned previously, DCMs provide a higher level of reliability for their estimates than comparable IRT or CTT models (Templin & Bradshaw, 2013)
  - It is easier to place a respondent into one of two groups (mastery or non-mastery) than to locate them on a scale

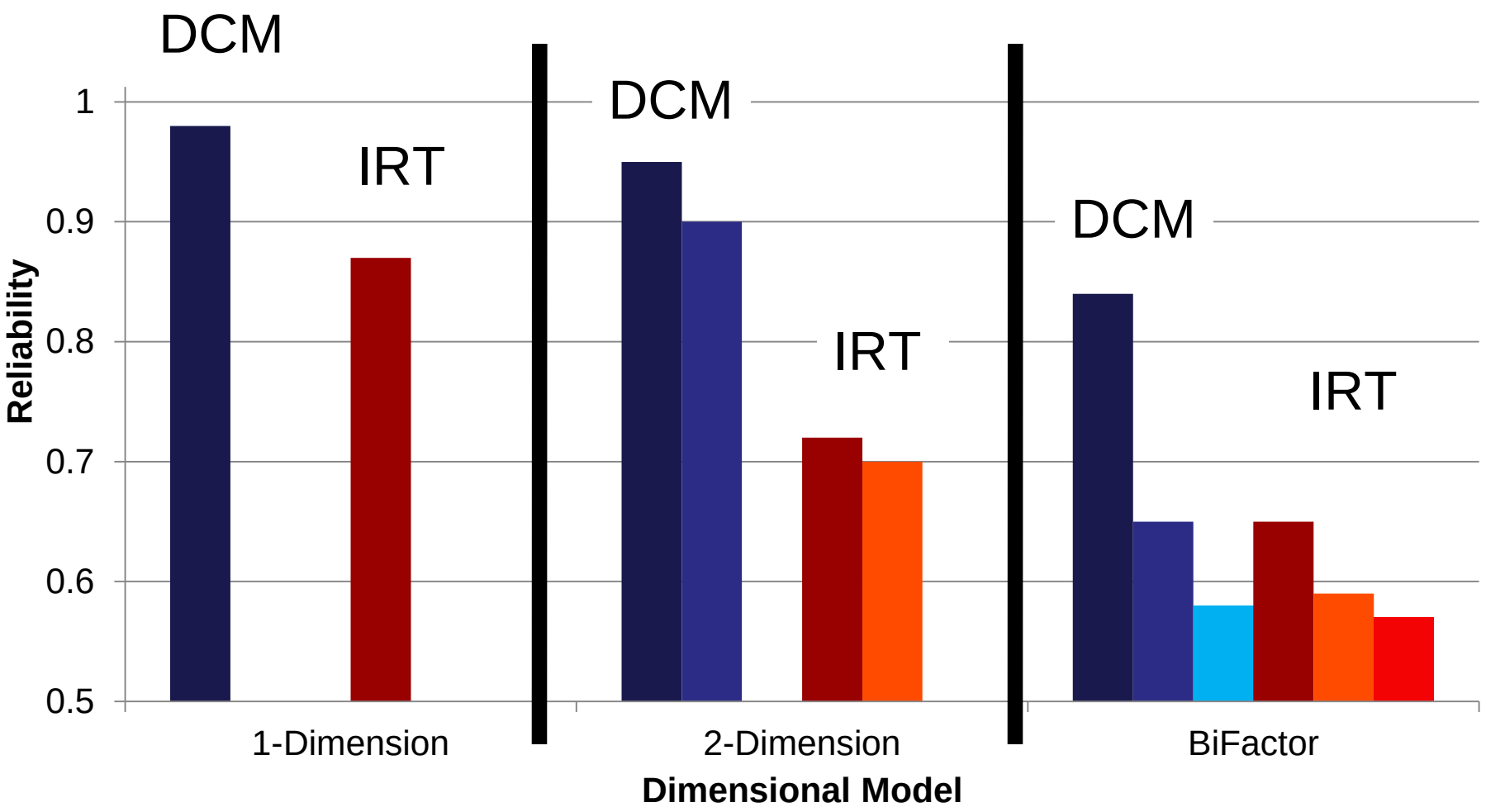
- Such characteristics allow DCMs to potentially change how large scale testing is conducted
  - Most EOC-type tests are for classification
    - Proficiency standards
  - DCMs provide direct link to classification
    - And direct access to standards

Theoretical Reliability Comparison

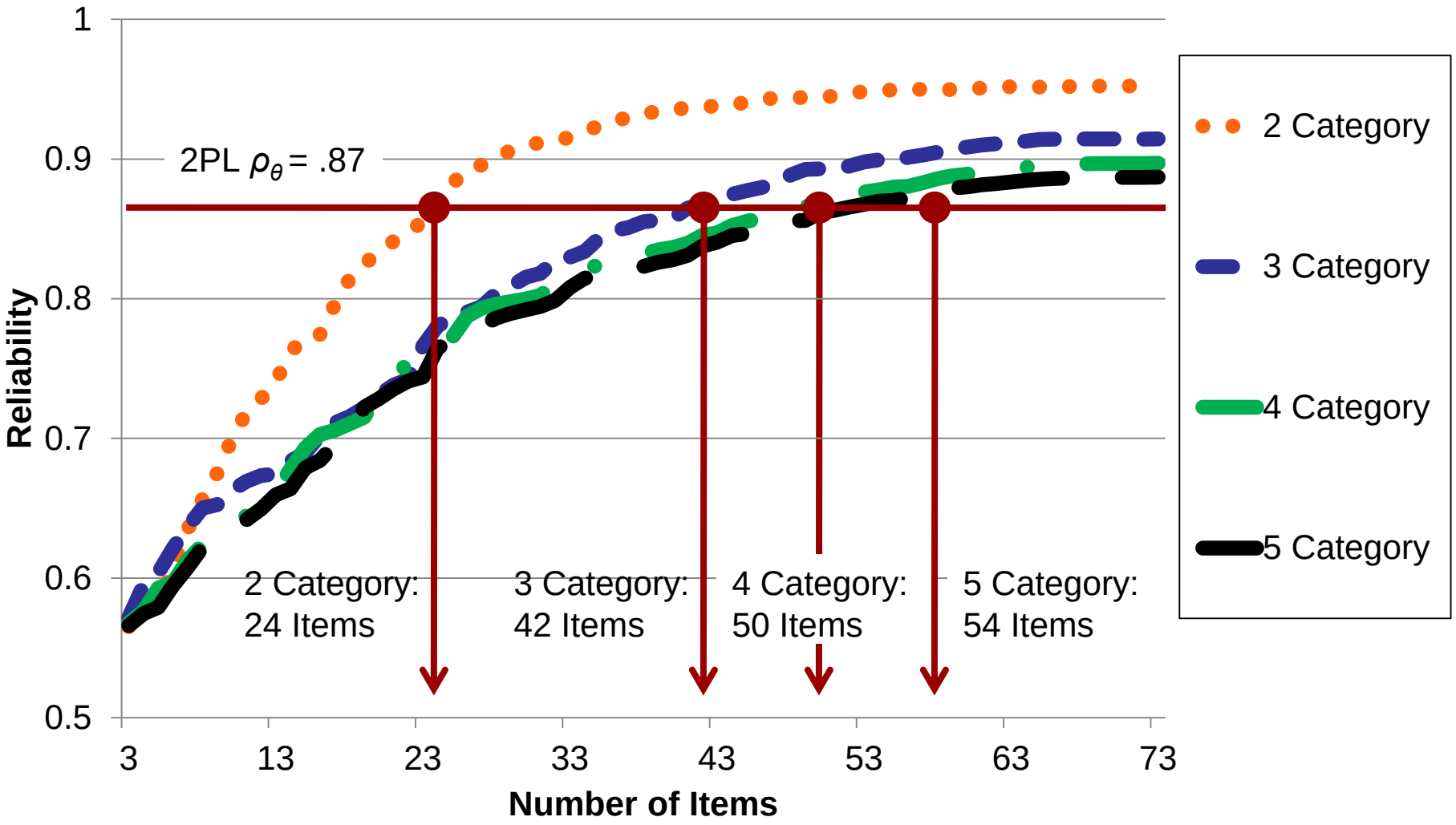


Reliability Level	DCM	IRT
0.80	8 Items	34 Items
0.85	10 Items	48 Items
0.90	13 Items	77 Items

Uni- and Multidimensional Comparison



DCMs for an EOC Test



Ramifications for Use of DCMs

- 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)
  - Multidimensional proficiency standards
    - Respondents must demonstrate proficiency on multiple areas to be considered proficient for an overall content domain
  - "Teaching to the test" would therefore represent covering more curricular content to best prepare respondents

- Shorter unidimensional tests
  - Unidimensional DCM application to empirical data:
    - Test needed only 24 items to have same reliability as IRT with 73 items

Ramifications for Use of DCMs: Formative Assessment

- Classroom appropriate test lengths
  - Teaching and testing time is limited
- Multivariate feedback
  - Strengths and weaknesses profiles
- No scores
  - Argued as a key element of effective formative testing in research

The Paradox of DCMs

- DCMs are often pitched as models that allow for measurement of "fine-grained" skills (e.g., Rupp & Templin, 2008)

- Paradox of DCMs:
  - Sacrifice fine-grained measurement of a latent trait for only several categories
  - Increased capacity to measure ability multidimensionally

When Are DCMs Appropriate?

- Which situations lend themselves more naturally to such diagnoses?
  - The *purpose* of the diagnostic assessment matters most
  - DCMs provide classifications directly
    - Optimally used when tests are used for classification
      - EOC Tests
      - Licensure/certification
      - Clinical screening
      - College entrance
      - Placement tests
  - DCMs *can* be used as approximations to continuous latent variable models
    - i.e., EOG example (2-5 category levels shown)

BENEFITS OF DCMS OVER TRADITIONAL CLASSIFICATION METHODS

Previous Methods for Classification

- Making diagnoses on the basis of test responses is not a new concept
  - Classical test theory
  - Item response theory
  - Factor analysis

- Process is a two-stage procedure
  1. Scale respondents
  2. Find appropriate cut-scores

- Classify respondents based on cut-scores

Problems with the Two-Stage Approach

- The two-stage procedure allows for multiple sources of error to affect the results

1. The latent variable scores themselves: estimation error
   - Uncertainty is typically not accounted for in the subsequent classification of respondents (i.e., standard errors)
   - The classification of respondents at different locations on the score continuum with multiple cut-scores is differentially precise
     - Uncertainty of the latent variable scores varies as a function of the location of the score

Problems with the Two-Stage Approach

2. Latent variable assumptions: that latent variable scores follow a continuous, typically normal, distribution
   - Estimates reflect the assumed distribution
   - Can introduce errors if the assumption is incorrect

3. Cut-score determination
   - Standard setting is imprecise when used with general abilities
     - Standard setting methods can be directed to item performance
   - Some theoretical justification needs to be provided for such a cut-off

Why are DCMs Better for Classification?

- The need for a two-stage procedure to set cut-scores for classification is eliminated when DCMs are used
  - Reduces classification error
- Quantifies and models the measurement error of the observable variables
  - Controlling for measurement error when producing the diagnosis
- DCMs have a natural and direct mechanism for incorporating base-rate information into the analysis
  - No direct way to do so objectively in two-stage procedures
- Item parameters provide information as to the diagnostic quality of each item
  - Not directly estimable in two-stage approaches
  - Can be used to build tests that optimally separate respondents

CONCLUDING REMARKS

Wrapping Up

- DCMs provide direct link between diagnosis and behavior
  - Provide diagnostic classifications directly
  - Diagnoses set by psychometric model parameters

- DCMs are effective if classification is the ultimate purpose
  - Reduce error by removing judgments necessary in two-stage approach

- DCMs can be used in many contexts
  - Can be used to create highly informative tests
  - Can be used to measure multiple dimensions