Diagnostic Measurement:
Theory, Methods, and Applications

Laine Bradshaw
University of Georgia

Jonathan Templin
University of Kansas

Course Website: http://wp.me/p3nkOf-nu
INTRODUCTIONS/COURSE MATERIALS: 8:00 – 8:15am

- **Section 1: Diagnostic Measurement Introduction** (8:15 – 9:45am)
  - Conceptually, how is “diagnostic” measurement different from more traditional measurement?
  - Why/when would you use diagnostic classification models (DCMs)?

  **BREAK: 9:45-10:00am**

- **Section 2: Theoretical Framework of DCMs** (10:00 – 11:45am)
  - Latent variables in diagnostic measurement
  - Specification of a general DCM, the log-linear cognitive diagnosis model (LCDM)

- **Questions/Discussion** (11:45-12:00 noon)

  **LUNCH: 12:00-1:00pm**

- **Section 3: DCMs in Practice** (1:00 – 2:30 pm)
  - Applying DCMs to a test of rational numbers
  - How-to for estimating DCMs with Mplus

  **BREAK: 2:30-2:45pm**

- **Section 4: Structural Model Specifications** (2:45 – 3:45pm)
  - Specifying structural model
  - Making alterations in Mplus

- **Section 5: Questions/Discussion** (3:45 – 4:45pm)
  - Discussions about our/your practical experiences with DCMs
  - Opportunity to ask about more specific content not covered in general introduction

**EVALUATE TRAINING SESSION: 4:45-5:00pm**
Conceptual Foundations of Diagnostic Measurement

Session 1
Session Overview

• Key definitions

• Conceptual examples

• Why diagnostic models should be used instead of traditional classification methods

• Concluding remarks
Session 1: Conceptual Foundations of Diagnostic Measurement

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

NCME 2014: Diagnostic Measurement Workshop
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 is likely to be the true state of reality
• 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 workshop:

  ➢ Educational Measurement
    • The competencies (skills) that a person has or has not mastered
      – Leads to possible tailored instruction and remediation

  ➢ Psychiatric Assessment
    • The DSM criteria that a person meets
      – Leads to a broader diagnosis of a disorder
**Workshop 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 categorical 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
Session 1: Conceptual Foundations of Diagnostic Measurement

CONCEPTUAL EXAMPLE
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
What do we learn from assessments?
Diagnostic Need in Education

• In education, there is a need for more specific feedback about what students do and do not understand

• Imagine if we could
  ➢ Diagnose mastery with respect to a set of skills
    • Complement standards-based curriculum (K-12)
    • Monitor students’ strengths and weaknesses
    • Place students into appropriate courses
  ➢ Tailor instruction to individual needs

• Concept is not contentious
  ➢ Common Core calls for diagnostic assessments
    • NCLB required diagnostic score reports
  ➢ Teachers seek more diagnostic score reports
TRADITIONAL MEASUREMENT MODELS
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.
Traditional testing procedures measure an **overall** ability in an area with a **continuous** latent variable.

Math ability is a latent variable.

The more math ability a person has the more likely he or she is to answer an item correctly.

Responses to items are observed variables.

**Math Ability**
- 12 - 4
- 4 + 4
- 6 / 3
- 2 * 7 + 12
Traditional Testing and Classification Methods

- **Information from Continuum:**
  - **Spencer** has more math ability than **Sue**
  - **Juan** scored in the 70th percentile
  - **Hugh** scored a 240 on the test
- **Diagnosis from Cut Score:**
  - **Sue** scored below the cut score
  - **Sue** will take the remedial math course
Traditional Psychometrics

• What results is a (weak) ordering of respondents
  ➢ Ordering is called weak because of error in estimates
  ➢ Hugh>Juan>Spencer>Sue

• Questions that traditional psychometrics cannot answer:
  ➢ Why is Sue so low?
    • How can we get her some help?
  ➢ How much ability is “enough” to pass?
    • How much is enough to be proficient?
  ➢ What math skills have the students mastered?
    • Which skills have they yet to master?
The item response is a function of a student’s ability.

The probability that Juan will answer this item correctly is .88.

The probability that Spencer will answer this item correctly is .52.
Multiple Dimensions of Ability

- As an alternative, we could have expressed math ability as a set of basic skills:
  - 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
DIAGNOSTIC MEASUREMENT MODELS
• Instead of measuring an overall math ability, “math” is expressed as a set of skills or *attributes*:
  - Add
  - Subtract
  - Multiply
  - Divide

• These attributes are categorical latent variables
  - Not a continuum

• Attributes are often dichotomous
  - Two levels or groups or categories
  - The two groups may be labeled differently:
    - Mastery/non-mastery
    - Proficient/Emerging
    - Ready/not ready
  - Mastery of an attribute ($\alpha = 1$) or non-mastery of an attribute ($\alpha = 0$)
What would our test look like?
- Each item measures one or more attributes

The attributes measured by each item are recorded in a **Q-matrix**
- Describes whether an item measures an attribute \((q = 1)\) or not \((q = 0)\)
- Mapping is established by content experts

### DCMs: A Conceptual Example

<table>
<thead>
<tr>
<th></th>
<th>Add</th>
<th>Subtract</th>
<th>Multiply</th>
<th>Divide</th>
</tr>
</thead>
<tbody>
<tr>
<td>12-4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4+4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6/3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2x7+12</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Diagnostic Classification Models

- DCMs uses responses to items to place students into **groups** according to **multiple** skills
  - No cut score is used to put students into groups; the model is built to do that

<table>
<thead>
<tr>
<th></th>
<th>Masters</th>
<th>Non-masters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add</td>
<td><img src="image" alt="Add Masters" /></td>
<td><img src="image" alt="Add Non-masters" /></td>
</tr>
<tr>
<td>Subtract</td>
<td><img src="image" alt="Subtract Masters" /></td>
<td><img src="image" alt="Subtract Non-masters" /></td>
</tr>
<tr>
<td>Multiply</td>
<td><img src="image" alt="Multiply Masters" /></td>
<td><img src="image" alt="Multiply Non-masters" /></td>
</tr>
<tr>
<td>Divide</td>
<td><img src="image" alt="Divide Masters" /></td>
<td><img src="image" alt="Divide Non-masters" /></td>
</tr>
</tbody>
</table>
Diagnostic Classification Models

- Students receive skill-specific feedback

<table>
<thead>
<tr>
<th></th>
<th>Add</th>
<th>Subtract</th>
<th>Multiply</th>
<th>Divide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spencer</td>
<td>![Green Check]</td>
<td>![Green Check]</td>
<td>![Green Check]</td>
<td>![Green Check]</td>
</tr>
<tr>
<td>Bill</td>
<td>![Green Check]</td>
<td>![Green Check]</td>
<td>![Green Check]</td>
<td>![Green X]</td>
</tr>
<tr>
<td>.84</td>
<td>.76</td>
<td>.35</td>
<td>.28</td>
<td></td>
</tr>
<tr>
<td>[Green Check]</td>
<td>![Green Check]</td>
<td>![Green Check]</td>
<td>![Green Check]</td>
<td></td>
</tr>
</tbody>
</table>

Spencer has mastered addition and subtraction, but should improve his ability to multiply and divide.

Or in Bill Gate-ese: Hey, Spencer you’ve got multiplication and division screwed up, but you’re fine on addition and subtraction.

- The model provides a probability each attribute is mastered
- Notice there is no “score” in a traditional or grading sense
The pattern of skills students have is often referred to as a mastery profile or an attribute profile.

- 4 patterns below are 4 of the possible 16 profiles of mastery that exist if there are 4 attributes:

<table>
<thead>
<tr>
<th></th>
<th>Add</th>
<th>Subtract</th>
<th>Multiply</th>
<th>Divide</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

All possible patterns:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Add</th>
<th>Sub.</th>
<th>Multiply</th>
<th>Divide</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

A classification of each individual skill results in a classification into one of these patterns:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Add</th>
<th>Sub.</th>
<th>Multiply</th>
<th>Divide</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Expected Examinee Responses

Given a student’s profile of mastery and the set of skills measured by each item, we can determine how we expect a student to respond to each item.

These students will likely answer the following items correctly:

- All Items
- Items 1, 2, 4
- Items 1, 2
- Item 2

Which items a student answers correctly depends on (1) which attributes he or she has and, (2) which attributes are being measured by the item.
- Example item: $2 \times 7 + 12 = ?$
  - Measures Attribute 1 (addition) and Attribute 3 (multiplication)

- Juan’s probability of correct response is .88
- Spencer’s probability of correct response is .3
**Traditional Score Report**

- Example from Georgia’s federally mandated end-of-course assessments

---

### Criterion-Referenced Competency Tests (CRCT) • Spring 2009

<table>
<thead>
<tr>
<th>Subject</th>
<th>Score</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading—GPS</td>
<td>800</td>
<td>Meets</td>
</tr>
<tr>
<td>English/Language Arts—GPS</td>
<td>790</td>
<td>Does Not Meet</td>
</tr>
<tr>
<td>Mathematics—GPS</td>
<td>870</td>
<td>Exceeds</td>
</tr>
<tr>
<td>Science—GPS</td>
<td>870</td>
<td>Exceeds</td>
</tr>
<tr>
<td>Social Studies—GPS</td>
<td>796</td>
<td>Does Not Meet</td>
</tr>
</tbody>
</table>

**Lexile: 650L**

Get to classification from score by using cut scores.

---

**Breakdown content area into sub-domains**

**Diagnostic Information**

**Sub-scores**

What can DCMs do differently?

- Example from Georgia’s federally mandated end-of-course assessments

DCMs can provide statistical probabilities an individual has mastered each sub-domain.

Get to classification from score by using cut scores.

Breakdown content area into sub-domains
# Diagnostic Score Report

**Student Name:** Daphne

## Review Your Answers

<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Your Answer</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>a</td>
<td>c</td>
<td>✓</td>
<td>d</td>
<td>✓</td>
<td>✓</td>
<td>c</td>
<td>✓</td>
<td>d</td>
<td>a</td>
<td>✓</td>
<td>b</td>
<td>a</td>
<td>d</td>
<td>c</td>
<td>b</td>
<td>a</td>
<td>c</td>
<td>b</td>
<td>a</td>
<td>c</td>
<td></td>
</tr>
<tr>
<td><strong>Correct Answer</strong></td>
<td>d</td>
<td>a</td>
<td>b</td>
<td>d</td>
<td>d</td>
<td>a</td>
<td>b</td>
<td>d</td>
<td>c</td>
<td>a</td>
<td>b</td>
<td>d</td>
<td>c</td>
<td>a</td>
<td>d</td>
<td>a</td>
<td>c</td>
<td>b</td>
<td>d</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>d</td>
<td>b</td>
</tr>
<tr>
<td><strong>Difficulty</strong></td>
<td>e</td>
<td>e</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>h</td>
<td>h</td>
<td>h</td>
<td>h</td>
<td>m</td>
<td>e</td>
<td>e</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>h</td>
<td>m</td>
<td>m</td>
<td>h</td>
<td>h</td>
<td>h</td>
<td>h</td>
<td>h</td>
</tr>
</tbody>
</table>

## 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

<table>
<thead>
<tr>
<th>Science Skill</th>
<th>Estimated Probability of Skill Mastery</th>
<th>Example Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systems</td>
<td>0.97</td>
<td>3, 14, 2, 17, 19, 23, 9</td>
</tr>
<tr>
<td>Classification</td>
<td>0.94</td>
<td>3, 12, 13, 5, 2, 17, 18, 16, 24, 7</td>
</tr>
<tr>
<td>Observation</td>
<td>0.45</td>
<td>11, 15, 1, 8, 18</td>
</tr>
<tr>
<td>Measurement</td>
<td>0.07</td>
<td>22, 20, 10, 11, 5, 6, 18, 25</td>
</tr>
<tr>
<td>Prediction</td>
<td>0.97</td>
<td>4, 14, 20, 12, 5, 19, 9</td>
</tr>
<tr>
<td>Data</td>
<td>0.74</td>
<td>22, 1, 19, 21</td>
</tr>
</tbody>
</table>

---

Diagnostic classification models (DCMs) are a set of statistical tools that provide diagnostic feedback.

Feedback is diagnostic in the sense that the decision itself—not just information to inform the decision—is given from the model.

No need for a two-stage approach of lining up and then cutting.
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
  - In Session #2 we discuss a general DCM
  - 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:

<table>
<thead>
<tr>
<th></th>
<th>Measurement Model</th>
<th>Structural Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Item/variable information (item parameters)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>How respondents with different diagnostic profiles perform on a set of test items</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Helps determine which items are better at discriminating between respondents with differing diagnostic profiles</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Respondent information pertaining to the base-rate or proportion of respondents with diagnoses in the population</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Provides frequency of diagnosis (or diagnostic profile)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Helps validate the plausibility of the observed diagnostic profiles</td>
<td></td>
</tr>
</tbody>
</table>
WHY USE DCMS OVER TRADITIONAL APPROACH?
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
  - Higher reliability than comparable IRT/MIRT models
  - Complex item structures possible
Data Demands for Reliability (Unidimensional)

<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>

8 Item Reliability

Data Demands for Reliability (Multidimensional)

- Can we measure multiple attributes reliably, too?

  - 18 item diagnostic test → diagnose mastery of 3 attributes
  - 34 item traditional test → scale unidimensional ability

• What about the accuracy of the classification:
  – As sample size decreases?
  – As dimensionality increases?

Classifying into Multiple Groups

OTHER CONCEPTUAL EXAMPLES OF DCMS
Unidimensional, Multicategory DCMs

- DCMs can be unidimensional and used for AYP-like classification purposes

<table>
<thead>
<tr>
<th>Math Ability</th>
<th>Advanced</th>
<th>Proficient</th>
<th>Basic</th>
</tr>
</thead>
<tbody>
<tr>
<td>4+4</td>
<td><img src="image1" alt="Green" /> <img src="image2" alt="Red" /></td>
<td><img src="image3" alt="Blue" /></td>
<td><img src="image4" alt="Pink" /></td>
</tr>
<tr>
<td>12/2</td>
<td><img src="image1" alt="Green" /> <img src="image2" alt="Red" /></td>
<td><img src="image3" alt="Blue" /></td>
<td><img src="image4" alt="Pink" /></td>
</tr>
<tr>
<td>16-3x4</td>
<td><img src="image1" alt="Green" /> <img src="image2" alt="Red" /></td>
<td><img src="image3" alt="Blue" /></td>
<td><img src="image4" alt="Pink" /></td>
</tr>
<tr>
<td>6/3</td>
<td><img src="image1" alt="Green" /> <img src="image2" alt="Red" /></td>
<td><img src="image3" alt="Blue" /></td>
<td><img src="image4" alt="Pink" /></td>
</tr>
</tbody>
</table>

- Math ability is unidimensional
- Math ability is categorical, not continuous

- The classification is statistical (directly from the model)
What can DCMs do differently?

- Example from Georgia’s federally mandated end-of-course assessments

<table>
<thead>
<tr>
<th>Criterion-Referenced Competency Tests (CRCT)</th>
<th>Spring 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name: JOHNSON, JENNY A</td>
<td></td>
</tr>
<tr>
<td>GTID: 5678123499</td>
<td></td>
</tr>
<tr>
<td>Gender: F</td>
<td></td>
</tr>
<tr>
<td>Grade: 5</td>
<td></td>
</tr>
<tr>
<td>Lexile: 650L</td>
<td></td>
</tr>
<tr>
<td>CONTENT AREA</td>
<td>SCORE</td>
</tr>
<tr>
<td>Reading-GPS</td>
<td>800</td>
</tr>
<tr>
<td>English/Language Arts-GPS</td>
<td>790</td>
</tr>
<tr>
<td>Mathematics-GPS</td>
<td>870</td>
</tr>
<tr>
<td>Science-GPS</td>
<td>870</td>
</tr>
<tr>
<td>Social Studies-GPS</td>
<td>796</td>
</tr>
</tbody>
</table>

Subjects can directly classify students into proficiency categories by assuming ability is unidimensional, but categorical.

IRT vs DCM Reliability

- DCMs yield equal reliability with fewer items (which means shorter tests)

Attributes are still categorical, but now have 3 categories instead of 2.

DCMs can place students into more than 2 groups.

<table>
<thead>
<tr>
<th></th>
<th>Advanced</th>
<th>Proficient</th>
<th>Basic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add</td>
<td><img src="image1" alt="Advanced" /></td>
<td><img src="image2" alt="Proficient" /></td>
<td><img src="image3" alt="Basic" /></td>
</tr>
<tr>
<td>Subtract</td>
<td><img src="image1" alt="Advanced" /></td>
<td><img src="image2" alt="Proficient" /></td>
<td><img src="image3" alt="Basic" /></td>
</tr>
<tr>
<td>Multiply</td>
<td><img src="image1" alt="Advanced" /></td>
<td><img src="image2" alt="Proficient" /></td>
<td><img src="image3" alt="Basic" /></td>
</tr>
<tr>
<td>Divide</td>
<td><img src="image1" alt="Advanced" /></td>
<td><img src="image2" alt="Proficient" /></td>
<td><img src="image3" alt="Basic" /></td>
</tr>
</tbody>
</table>
Session 1: Conceptual Foundations of Diagnostic Measurement

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
Ramifications for Use of DCMs

• Reliable measurement of multiple dimensions is possible
  ➢ 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 diagnosis?

➢ 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 coarse approximations to continuous latent variable models
  • i.e., EOC example (2-5 category levels shown)
Session 1: Conceptual Foundations of Diagnostic Measurement

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
Session 1: Conceptual Foundations of Diagnostic Measurement

CONCLUDING REMARKS
Session 1 – Take-home Points

- 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