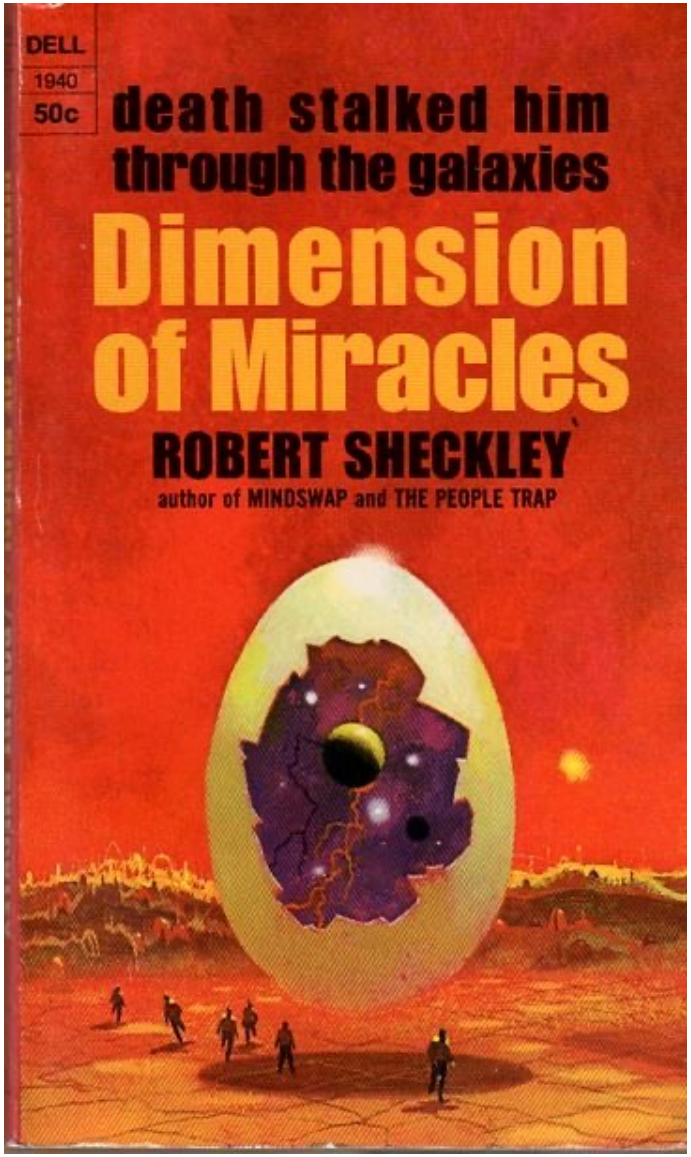
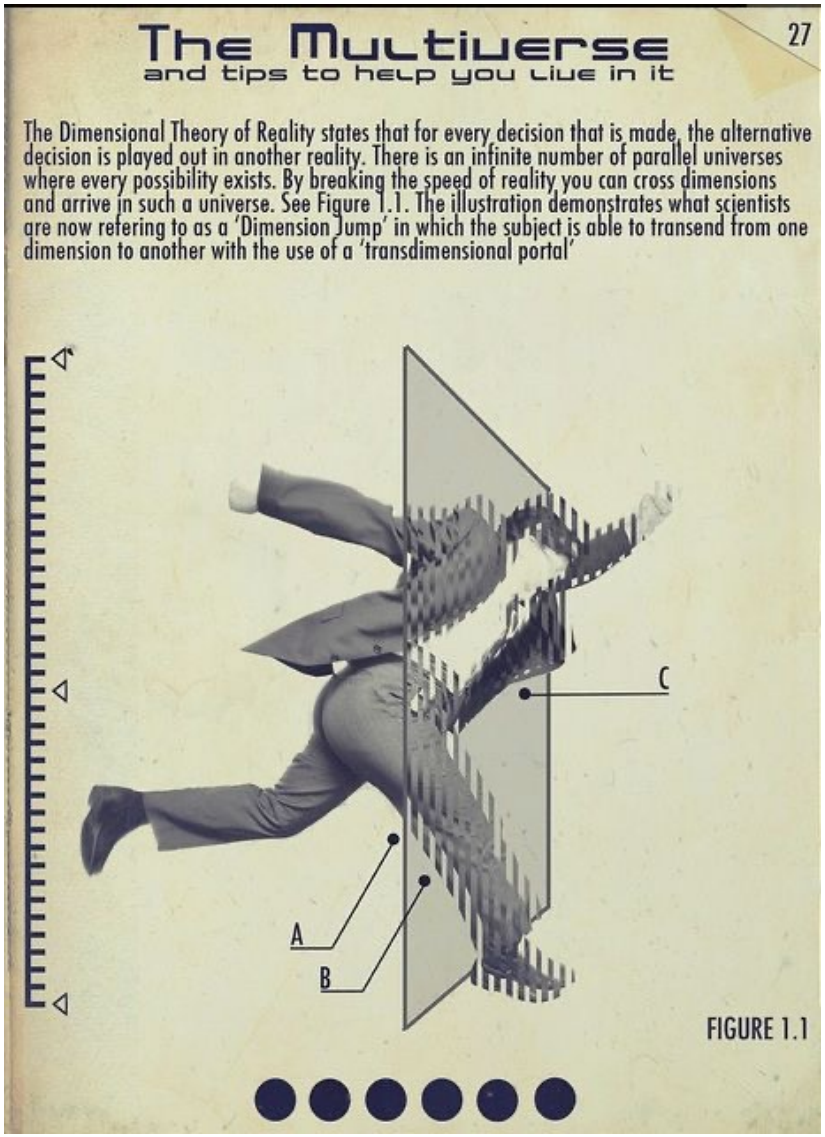
IOWA

MMM 2023: Week 01

Tales from an Alternate Dimension: Customs and Practices in Multidimensional Assessment Methods

Jonathan Templin


DELL
1940
50c
death stalked him through the galaxies
Dimension of Miracles
ROBERT SHECKLEY
author of MINDSWAP and THE PEOPLE TRAP


The Multiverse
and tips to help you live in it
27
The Dimensional Theory of Reality states that for every decision that is made, the alternative decision is played out in another reality. There is an infinite number of parallel universes where every possibility exists. By breaking the speed of reality you can cross dimensions and arrive in such a universe. See Figure 1.1. The illustration demonstrates what scientists are now refering to as a 'Dimension Jump' in which the subject is able to transend from one dimension to another with the use of a 'transdimensional portal'
C
A
B
FIGURE 1.1

Talk Overview

- Framework: Generalized latent variable models
  - Measurement models
  - Structural models
- Multidimensionality: feature or a bug?
- Can multidimensionality be a good thing?
- Not all tests are multidimensional
  - Beware of subscores from unidimensional assessments

Key Definitions

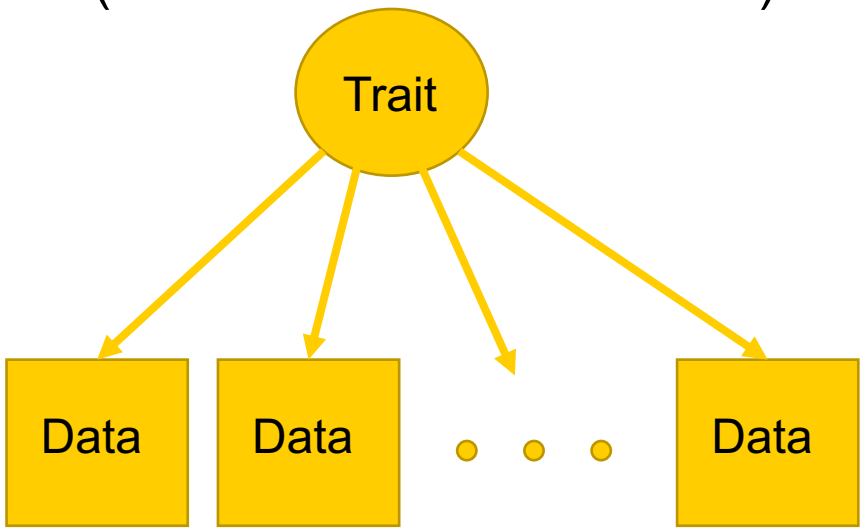
- **Score(s):** numeric representations of some type of behavior
  - Latent trait(s), latent construct(s), latent variable(s)
  - Ability(ies)
  - Skill(s)
  - Attribute(s)
  - Dimension(s)
- **Assessment/test:** A collection of tasks used to quantify scores
- **Item:** A generic term for an assessment task
- **Person/student/examinee/respondent:** The person giving you the data
- **Psychometric Models/Methods:** The mathematical linkage between a person's data and their latent trait(s)

Psychometric Models

Without putting labels on it, the core of a psychometric model can be represented by a very simple equation

$$Data \leftarrow Trait(s)$$

Or, a path diagram (shown with one latent trait):



Small Sidebar

- When considering psychometric models, envision the latent variables as being observed

- Why?
  - Linkages to linear models
    - CFA: multiple regression
    - IRT: logistic regression
    - MIRT: logistic regression with multiple predictors
    - DCMs: ANOVA
  - Design of assessments/administration of assessments follows experimental and observational design fundamentals

- Examples
  - Unidimensional model (IV predicts DV)
  - Multidimensional model (IV predicts DV controlling for other IVs)

Items: The Data Side

- Items are the observed data in psychometric models
- Item data come in many types:
  - Discrete categories
    - Binary/dichotomous responses
    - Multicategory/Polytomous responses (ordered or unordered)
    - Continuous responses (e.g., reaction time, slider bar data)
    - Counts
- For today, we'll have the following details
  - A response to item i by person p is Y_{pi}
- Why use Y when many psychometric texts use X?
  - Underscoring ties to linear models – data are the outcome, not the predictor

Theory: The Model Side

- The theory/model side is where the latent traits go
- Generically, we will denote this side using a regression-like set of parameters

$$\beta_{i0} + \beta_{i1}\theta_{p1} + \cdots + \beta_{iD}\theta_{pD}$$

- Here, the latent variables are denoted as $\theta_{p1}, \ldots, \theta_{pD}$
  - These are placeholders—there likely won't be a case where all dimensions are measured by an item (despite what EFA suggests)

- The "item parameters" or measurement model parameters are (for now) denoted by the regression coefficients
  - β_{i0} → intercept for item i
  - $\beta_{i1}, \ldots, \beta_{iD}$ → slopes/factor loadings/item discriminations for item i for each dimension

Generalized Models

- I've not put data and theory in the same equation yet as the equation will take many different forms

- Generalized linear models are the overarching family by which models are often formed
  - Data → Expected value of item responses → link function → model
    - Enables "regression"-like model kernel (linear predictor)

- Example (single latent variable; assume mean zero and variance one):
  - Confirmatory factor analysis (identity link): $Y_{pi} = \mu_i + \lambda_i \theta_p + e_i$

$$g\left(E\left(Y_{pi}|\theta_p\right)\right) = 1 * E\left(Y_{pi}|\theta_p\right) = \mu_i + \lambda_i \theta_p$$

  - Item response theory (binary data/logit and other link functions):

$$g\left(E\left(Y_{pi}|\theta_p\right)\right) = \log\left(\frac{P\left(Y_{pi}=1|\theta_p\right)}{1-P\left(Y_{pi}=1|\theta_p\right)}\right) = \mu_i + \lambda_i \theta_p = a_i\left(\theta_p - b_i\right)$$

Where $a_i = \lambda_i$ and $b_i = -\frac{\mu_i}{\lambda_i}$

What's in a Model

- A psychometric model is a mathematical representation of students' item responses
  - Left side: Data
  - Right side: Student score(s), item properties, and design controls
    - Common item stimulus (e.g., testlets)
    - Item position
    - Speededness
    - Classroom/school effects (hierarchical/multilevel data)
- Many different options exist for what appears on each side
- Each side can be tailored to match the characteristics of a test or item (or other considerations)

If Latent Variables were Observed

- If latent variables were observed, how would we interpret the following items from unidimensional models:

$Y_{pi} = \mu_i + \lambda_i \theta_p + e_i$ (CFA)

$P(Y_{pi} = 1|\theta_p) = \frac{\exp(\mu_i + \lambda_i \theta_p)}{1 + \exp(\mu_i + \lambda_i \theta_p)}$ (IRT)

- What about the following multidimensional items?

$Y_{pi} = \mu_i + \lambda_{i1} \theta_{p1} + \lambda_{i2} \theta_{p2} + e_i$ (CFA)

$P(Y_{pi} = 1|\theta_p) = \frac{\exp(\mu_i + \lambda_{i1} \theta_{p1} + \lambda_{i2} \theta_{p2})}{1 + \exp(\mu_i + \lambda_{i1} \theta_{p1} + \lambda_{i2} \theta_{p2})}$ (IRT)

- What does that say about traits in models such as the bifactor model (or “method” factors in CFA)?

From Models to Scores

- The right-hand side of the model is where students' latent trait(s) are embedded
- As shown in the model equations, models use these traits to predict how students will respond to items
  - This seems to imply we have to know students' scores before using a model
- We determine scores through optimization functions involving likelihoods
  - Either Bayes' Theorem or Maximum Likelihood
- The model must agree with the data to ensure the scores (and their conditional standard errors/reliabilities) are accurate

Is Multidimensionality a Good Thing?

A Feature? A Bug? Or…and Entirely Different Question?

Multidimensionality: Feature or Bug?

- A surprising question that needs asked when considering multidimensional psychometric models is:
  - Is multidimensionality good or bad?
  - **Does your data match your assumptions about underling trait(s)**
- If assuming a single/unidimensional trait:
  - Multidimensionality is seen as bad
    - Need to determine if it is present and, if so, what to do about it
- If assuming multiple traits:
  - Multidimensionality is seen as good
    - Need to verify it exists “meaningfully”

Traditionally: Dimensionality Varies by Field

- Historically, those in psychology have been interested in multidimensional assessment
  - Example Theories: 16 PF, “big five”, Holland’s Vocational Interests Theory (six dimensions)
  - Often seek to “verify” structure of items
  - Interested in pattern of latent trait correlations (see Rounds’ modifications of Holland’s Theory onto a circumplex)

- Unidimensional assessment is most common in educational measurement applications of measurement models
  - So, often multidimensionality is a bad thing!

When Multiple Dimensions are not Wanted

Methods for Determining if They Exist

MMM 2023: Week 01

When Multidimensionality Is Bad

- Unwanted multidimensionality?
  - Determine if it is present
  - Remove it (if possible)

- Applies to many different educational measurement situations
  - Almost all unidimensional tests

- Historical problem:
  - Many methods used for decades aren't as applicable with modern methods (e.g., exploratory factor analysis)

Investigating Multidimensionality

- Goal: Determine if multiple dimensions are present

- Historically: Exploratory Factor Analysis (EFA)
  - History of EFA is long…and much of it is prior to invention of computers or coding systems

- Key question: If you believe there is one dimension and have one model, why use other methods to evaluate this question?
  - Different results may be due to confound of different methods with different assumptions

The Many Flavors of EFA

- → Matrix decomposition-based (limited information) EFA
  - Input → Covariance (correlation) matrix
  - Output→
    - Eigenvalues (determine proportion of variance per dimension)
    - Eigenvectors (not needed in question of if multiple dimensions are present)
  - Problem: Matrix decomposition techniques are highly sensitive to sample characteristics and often give incorrect answers
    - Additionally: Which matrix do you decompose, Pearson or tetrachoric?
    - Further: Missing data assumptions are weakest (Missing Completely at Random)

- → Likelihood-based (full information) EFA
  - Better choice:
    - More robust missing data assumptions
    - Direct modeling of data
  - But #1: Must align with data type (binary items—EFA assuming binary items)
  - But #2: **Likelihood-based EFA does not exist (all are confirmatory analyses)**

The Syntax of Factor Analysis

- Factor analysis works by hypothesizing that a set of latent factors helps to determine a person's response to a set of variables
  - This can be explained by a system of simultaneous linear models

$$Y_{p1} = \mu_1 + \lambda_{11}\theta_{p1} + \lambda_{12}\theta_{p2} + \cdots + \lambda_{1D}\theta_{pD} + e_{p1}$$
$$Y_{p2} = \mu_2 + \lambda_{21}\theta_{p1} + \lambda_{22}\theta_{p2} + \cdots + \lambda_{2D}\theta_{pD} + e_{p2}$$
$$\vdots$$
$$Y_{pI} = \mu_I + \lambda_{I1}\theta_{p1} + \lambda_{I2}\theta_{p2} + \cdots + \lambda_{ID}\theta_{pD} + e_{pI}$$

- μ_i = intercept (mean in FA if latent means are zero) for variable i
- λ_{id} = factor loading for item i onto dimension d (regression slope/discrimination)
  - Factors are assumed distributed MVN with zero mean and (for EFA only) identity covariance matrix (uncorrelated factors – to start)
- e_{pi} = residual for person p and item i (only for continuous data)
  - Residuals are assumed distributed MVN (across items) with a zero mean and a diagonal covariance matrix $\boldsymbol{\Psi}$ containing the unique variances
- Often, this gets shortened into matrix form:

$$\mathbf{Y}_p = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\Theta}_p^T + \mathbf{e}_p$$

How Maximum Likelihood EFA Works

- Maximum likelihood EFA assumes the data follow a multivariate normal distribution
  - The basis for the log-likelihood function (same log-likelihood we have used in every analysis to this point)

- The log-likelihood function depends on two sets of parameters: the mean vector and the covariance matrix
  - Mean vector is saturated (just uses the item means for item intercepts) – so it is often not thought of in analysis

  - Covariance matrix is what gives “factor structure”
    - EFA models provide a structure for the covariance matrix

EFA Model-Implied Covariance Matrix

- The covariance matrix is modeled based on how it would look if a set of hypothetical (latent) factors had caused the data

- For an analysis measuring D dimensions, each item in the EFA:
  - Has 1 unique variance parameter
  - Has D factor loadings

- The initial estimation of factor loadings is conducted based on the assumption of uncorrelated factors
  - Assumption is dubious at best – yet is the cornerstone of the analysis

Model Implied Covariance Matrix

- The factor model implied covariance matrix is $\boldsymbol{\Sigma}_Y = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$
  - Where:
    - $\boldsymbol{\Sigma}_Y$ = model implied covariance matrix of the observed data (size I x I)
    - $\boldsymbol{\Lambda}$ = matrix of factor loadings (size I x D)
      - In EFA: all terms in $\boldsymbol{\Lambda}$ are estimated
    - $\boldsymbol{\Phi}$ = factor covariance matrix (size D x D)
      - In EFA: $\boldsymbol{\Phi} = \mathbf{I}$ (all factors have variances of 1 and covariances of 0)
      - In CFA: this is estimated
    - $\boldsymbol{\Psi}$ = matrix of unique (residual) variances (size I x I)
      - In EFA: $\boldsymbol{\Psi}$ is diagonal by default (no residual covariances)

- Therefore, the EFA model-implied covariance matrix is:

$$\boldsymbol{\Sigma}_Y = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$$

EFA Model Identifiability

- With ML EFA there are two rules for identification
  - T-rule: Total number of EFA model parameters must not exceed unique elements in saturated covariance matrix of data
    - For an analysis with D dimensions and I items there are $D^*I + D = I(D+1)$ EFA model parameters
    - As we will see, there must be $\frac{D(D-1)}{2}$ constraints for the model to be identified
    - Therefore, $I(D+1) - \frac{D(D-1)}{2} < \frac{I(I+1)}{2}$
  - Local-identification: each portion of the model must be locally identified
    - With all factor loadings estimated local identification fails
      - No way of differentiating factors without constraints

Constraints to Make EFA in ML Identified

- The EFA model imposes the following constraint:

$$\boldsymbol{\Lambda}^T \boldsymbol{\Psi} \boldsymbol{\Lambda} = \boldsymbol{\Delta}$$

such that $\boldsymbol{\Delta}$ is a diagonal matrix (see Lawley and Maxwell, 1972)

- This puts $\frac{D(D-1)}{2}$ constraints on the model (that many fewer parameters to estimate)

- This constraint is not well known – and how it functions is hard to describe
  - For a 1-factor model, the results of EFA and CFA will match

- Note: the other methods of EFA "extraction" avoid this constraint by not being statistical models in the first place
  - PCA-based routines rely on matrix properties to resolve identification

The Nature of the Constraints in EFA

- The EFA constraints provide some detailed assumptions about the nature of the factor model and how it pertains to the data

- For example, take a 2-factor model (one constraint):

$$\sum_{i=1}^{I} \psi_i^2 \prod_{d=1}^{2} \lambda_{id} = 0$$

- In short, some combinations of factor loadings and unique variances (across and within items) cannot happen
  - This goes against most of our statistical constraints – which must be justifiable and understandable (therefore testable)
  - This constraint is not testable in CFA

An Equivalent Model

- With EFA via ML CFA, the crazy constraint system employed on the previous slide can be avoided by using an equivalent model
  - Equivalent = same model loglikelihood and number of parameters

- How? Set matrix of loadings so that it is in "row-echelon" form
  - Set factor loadings to zero (or some other constant) for small set of items across some dimensions

Two Dimensions	Three Dimensions
$\begin{matrix} \lambda_{11} & 0 \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \end{matrix}$	$\begin{matrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \end{matrix}$

EFA for IRT: Constraints Needed

- When considering IRT, the key difference is that items are assumed to follow a Bernoulli distribution
  - No unique variance parameters
  - Different likelihood function

- Identification rules:
  - Total number of parameters is no longer governed by number of parameters in mean vector/covariance matrix
    - No longer multivariate normal likelihood
  - But: Local identification rules are maintained
    - Meaning: $\frac{D(D-1)}{2}$ constraints are needed

Model-based Dimensionality: Likelihood Ratio Tests and Approximate Fit Indices

- So, the modern, model-based methods for checking dimensionality don't use CFA or EFA
  - They use IRT and MIRT

- Using (Marginal) Maximum Likelihood:
  - Step 1: Estimate unidimensional IRT model
  - Step 2: Estimate two-dimensional IRT model with one discrimination set to constant (does not matter which one)
  - Step 3: Use likelihood ratio test of 1- vs. 2- dimensional model
    - If p-value is significant: multidimensional model fits better (but, in educational assessment can have very large sample sizes and massive power)
    - Check on limited information goodness of fit indices (using M2)

Motivations for Multiple Dimensions in Education

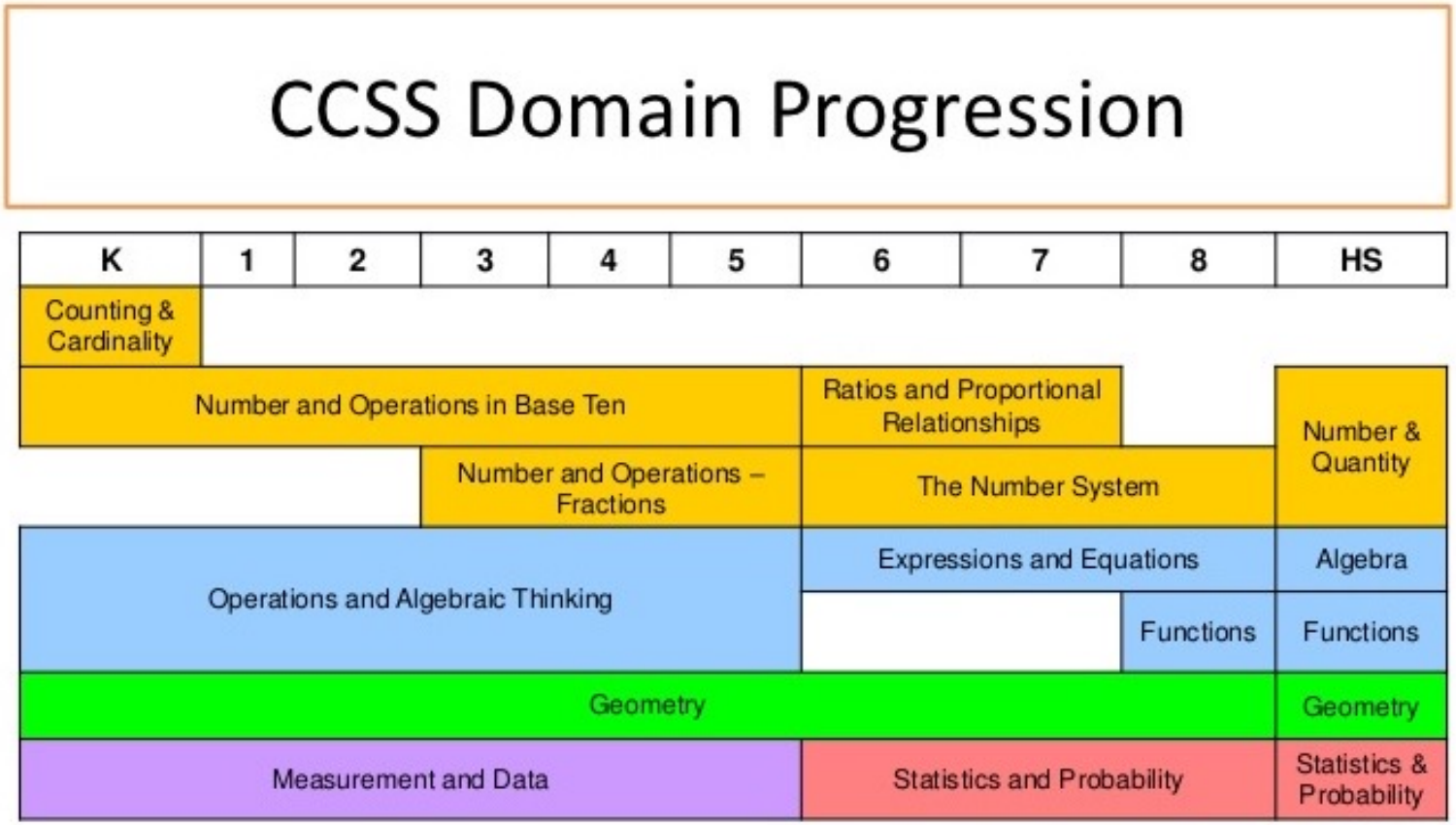
What if Multiple Dimensions Were Good Things?

MMM 2023: Week 01

Motivations for Multiple Dimensions

- There are good reasons for multidimensional assessment, *particularly* in educational measurement:
  - Additional information needed about student abilities
    - Profile of scores
    - Score reports with subscores
  - Some items measure/necessitate more than one dimension
    - Stay tuned for an example from a general math assessment
  - Some constructs/standards are inherently multidimensional
    - Curricular standards
    - Content domains
    - Test blueprints???

Types of Additional Information



https://www.slideshare.net/MarciShepard/common-core-state-standards-math-workgroup-training

Multidimensional Item Features

- Example: One reading passage with multiple questions

- Typical (unidimensional) 1PL IRT model:

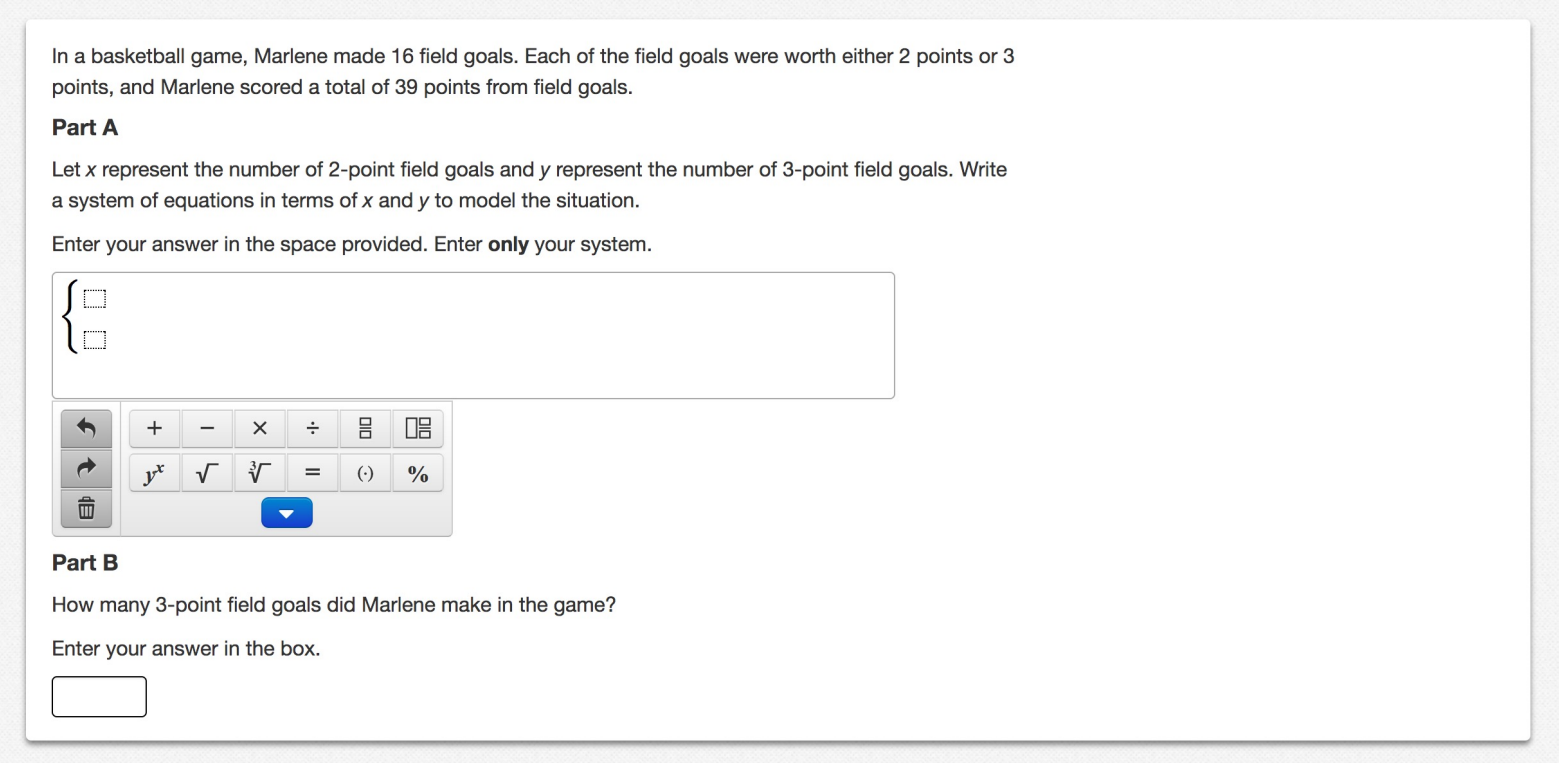
$$P\left(Y_{pi} = 1 \mid \theta_p\right) = \frac{\exp\left(\theta_p - b_i\right)}{1 + \exp\left(\theta_p - b_i\right)}$$

- Basic Testlet IRT model:

$$P\left(Y_{pi} = 1 \mid \theta_p\right) = \frac{\exp\left(\theta_p - b_i + \gamma_p\right)}{1 + \exp\left(\theta_p - b_i + \gamma_p\right)}$$

Item Content is Multidimensional

- An example from the PARCC practice test (Algebra 1):



In a basketball game, Marlene made 16 field goals. Each of the field goals were worth either 2 points or 3 points, and Marlene scored a total of 39 points from field goals.

Part A

Let x represent the number of 2-point field goals and y represent the number of 3-point field goals. Write a system of equations in terms of x and y to model the situation.

Enter your answer in the space provided. Enter **only** your system.

Part B

How many 3-point field goals did Marlene make in the game?

Enter your answer in the box.

Is Multidimensionality Good or Bad?

- Is being able to read needed to answer this item?

- If reading comprehension is needed for this item, is it measured by the model?
  - If not, bad for measurement of math ability

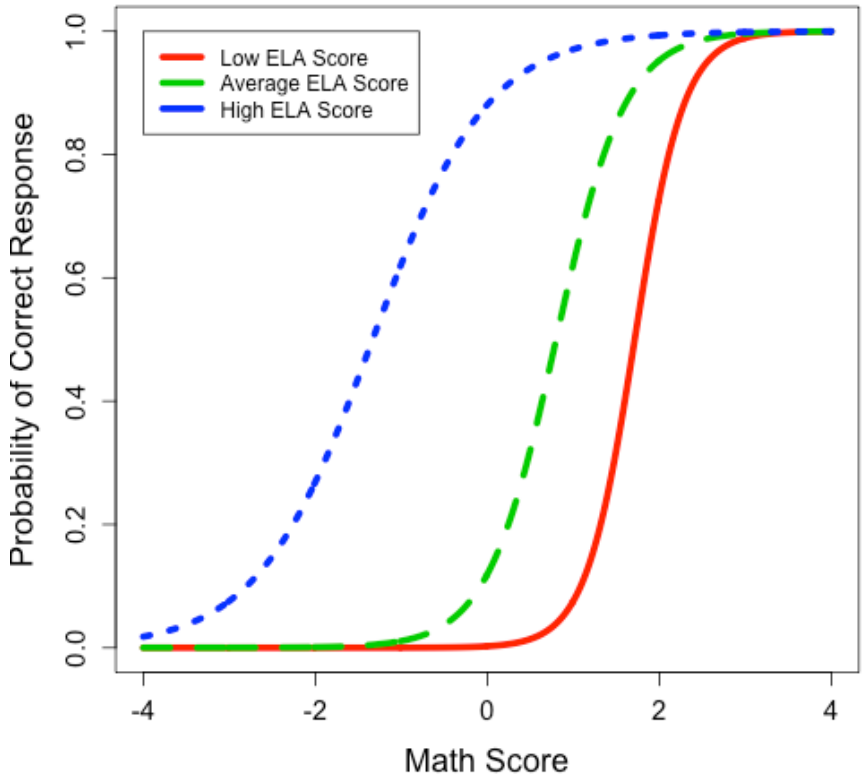
- A multidimensional IRT model is needed (Reading and Math):

$$P\left(Y_{pi}=1 \mid \theta_{pR},\theta_{pM}\right)=\frac{\exp\left(\beta_{i0}+\beta_{iR}\theta_{pR}+\beta_{iM}\theta_{pM}\right)}{1+\exp\left(\beta_{i0}+\beta_{iR}\theta_{pR}+\beta_{iM}\theta_{pM}\right)}$$

- (not typically asked) Does math and reading ability interact?

$$P\left(Y_{pi}=1 \mid \theta_{pR},\theta_{pM}\right)=\frac{\exp\left(\beta_{i0}+\beta_{iR}\theta_{pR}+\beta_{iM}\theta_{pM}+\beta_{iRM}\theta_{pR}\theta_{pM}\right)}{1+\exp\left(\beta_{i0}+\beta_{iR}\theta_{pR}+\beta_{iM}\theta_{pM}+\beta_{iRM}\theta_{pR}\theta_{pM}\right)}$$

What a MIRT Model with Latent Variable Interaction Looks Like



Benefits and Drawbacks to Multiple Dimensions

Dimensions Cannot Be Created: They Must Exist

Benefits of Multidimensionality

- Surprisingly, when assessed simultaneously, latent traits "borrow" information from the other, correlated, dimensions

- From a multivariate modeling perspective, the following occurs:
  - Latent traits have a direct effect on items they measure
    - And items directly provide information about the traits measured by them
  - Latent traits have indirect effects on items they do not measure but are measured by other latent traits with which they correlate
    - Additional information provided by these items

- In practice, this means that precision for the traits measured can increase when measuring them simultaneously

Example of Higher Precision

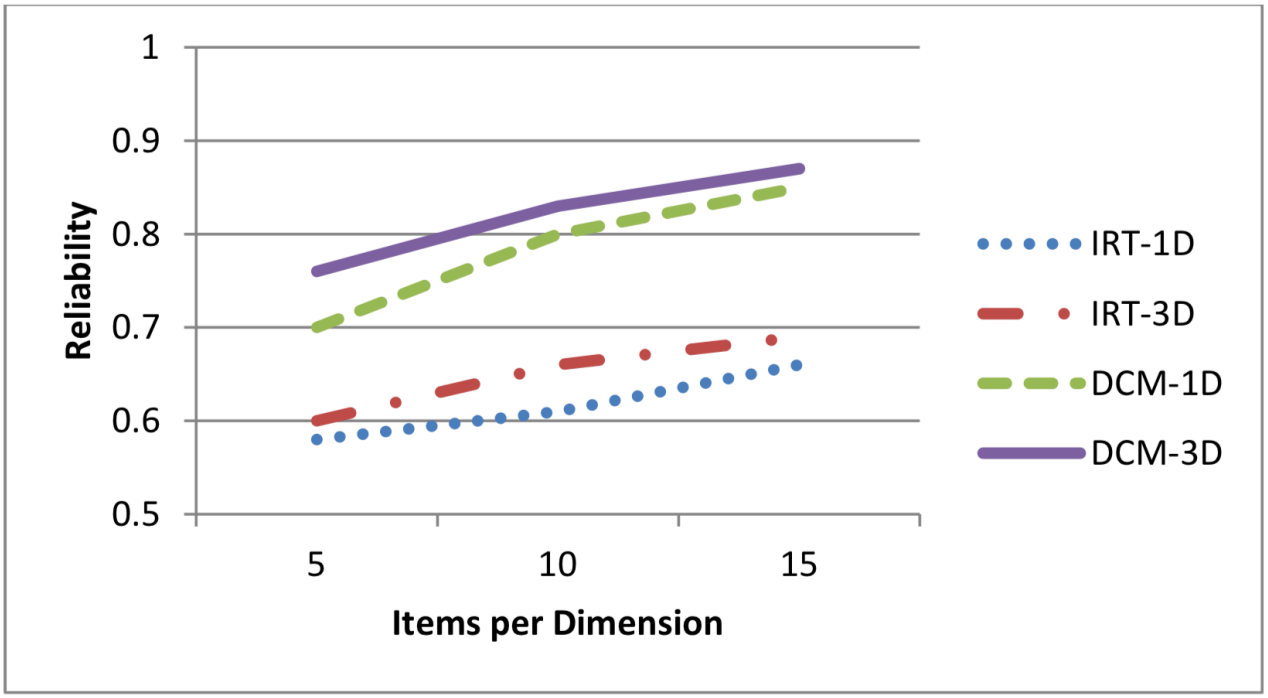


Figure 2. Simulation Study Results: Reliability of DCM and IRT Models with 100 Generated Tests

Source: Templin & Bradshaw (2013, Journal of Classification)

Drawbacks to Multidimensional Assessment

- Challenges that can be overcome:
  - In general, many more items overall are needed (think of this as items per dimension)
  - Items measuring multiple traits may have less information per trait (especially for categorical items)
  - Estimation is a challenge when not using CFA
    - Multidimensional integration is needed—time for accurate estimation increases exponentially with number of dimensions assessed
- Challenges that are difficult to overcome
  - In education, dimensionality is likely related to age, learning, and other person- and group-based characteristics
  - Dimensions cannot be created—they must exist meaningfully

Dimensions Must Exist

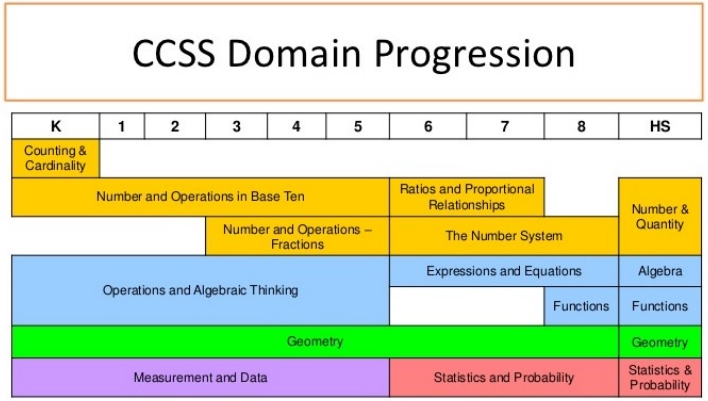
- Dimensions cannot be created—they must exist meaningfully

- This is indexed by examining the correlations of the factors
  - Very high correlations (e.g., greater than .9) indicate the same information is being provided by different dimensions

- Can demonstrate with a common educational assessment example: Scale subscores

Scale Subscores

- Consider a situation where an assessment program is tasked with, from a single test, providing:
  - A unidimensional score (e.g., mathematics ability)
  - A set of scale subscores (e.g., domain-level abilities)

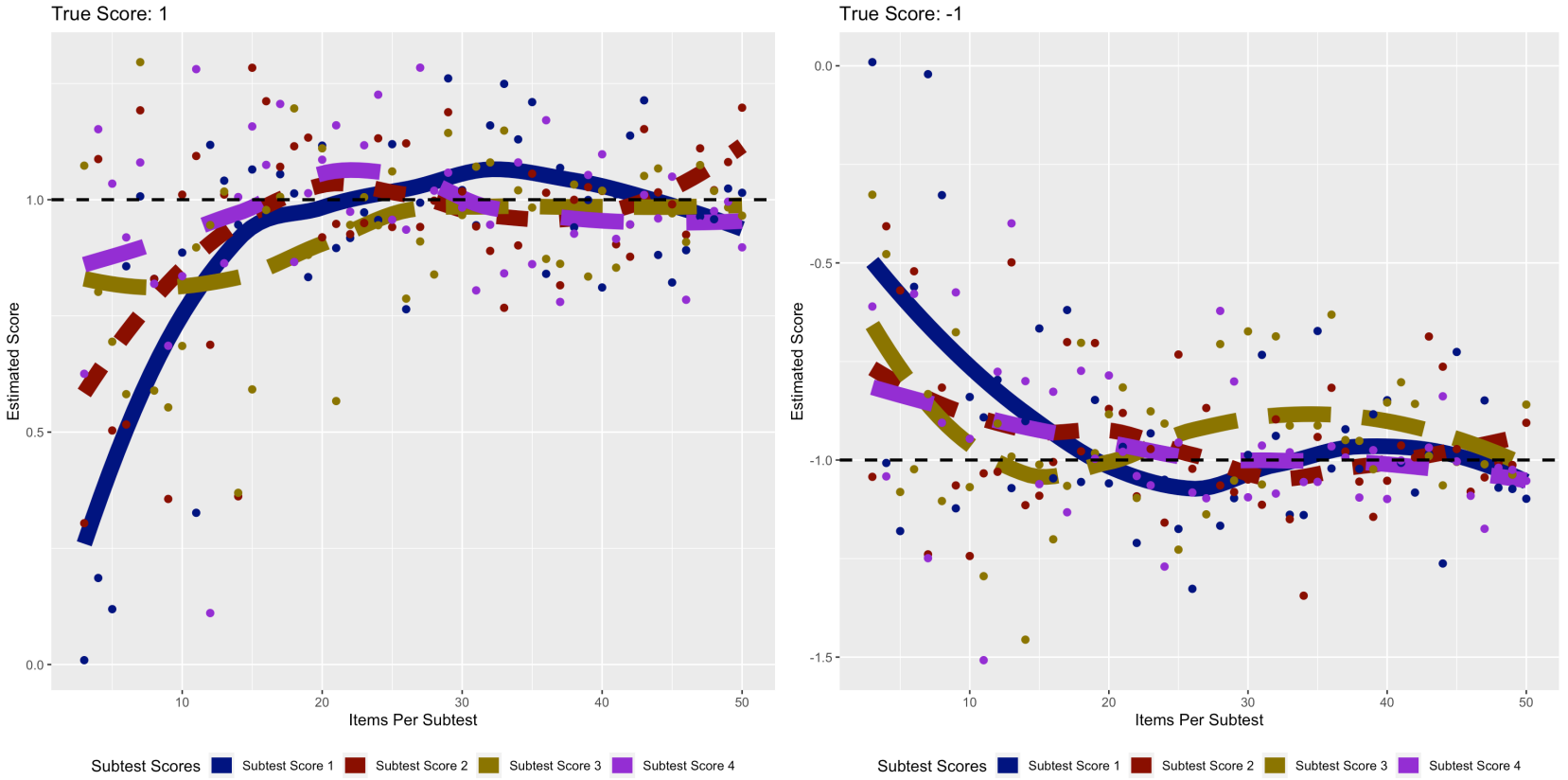


- The problem?
  - The unidimensional model and the multidimensional subscores cannot both be the correct model

- Typically, the solution is to focus on the overall score (the unidimensional model)
  - Assuming this is true, this makes subscores effectively smaller measures of the same unidimensional trait
    - With less precision
    - Correlated at one

- Any differences are from error!

Subscore Simulation



Wrapping Up

Concluding Remarks

- Building multidimensional assessments is difficult!
  - Theory must align with empirical evidence
  - Estimation must cooperate

- Today's talk was a collection of thoughts based on conversations (and professional disagreements) that occur frequently in my career

- Overall, the state of the field of multidimensional assessment remains siloed—but hopefully will change soon!

Thank you!

Questions? jonathan-templin@uiowa.edu