Blending Psychometrics with Bayesian Inference Networks: Measuring Hundreds of Latent Variables Simultaneously

Jonathan Templin
Department of Educational Psychology
Achievement and Assessment Institute
University of Kansas

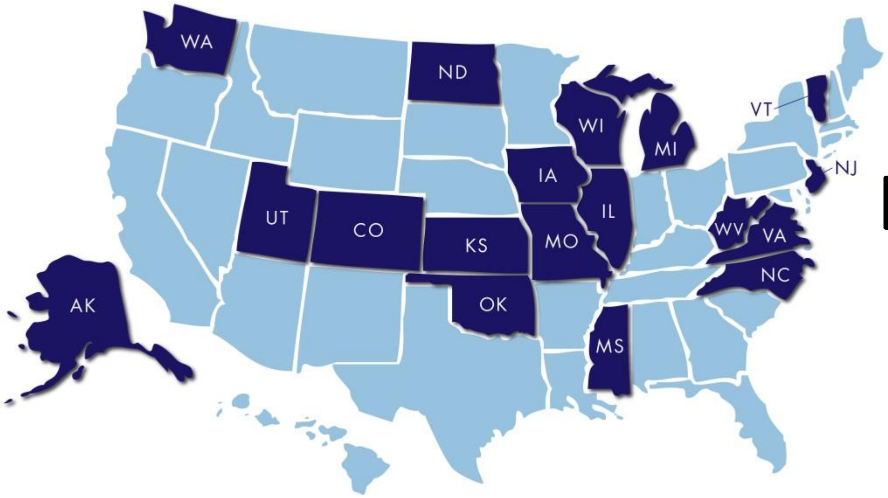
THE UNIVERSITY OF KANSAS

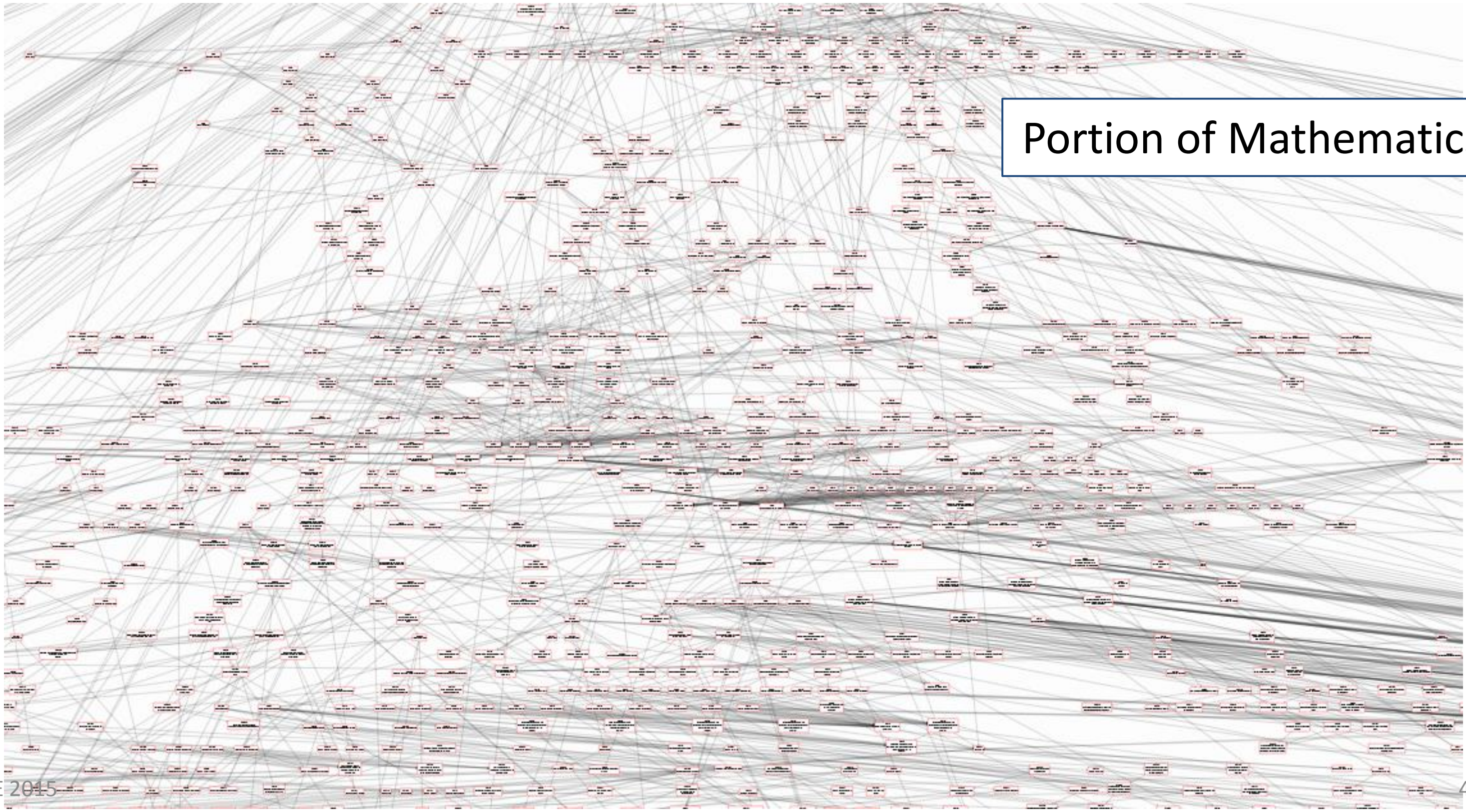
Today's Talk

- The issue: psychometric modeling for the Dynamic Learning Maps (DLM) project
  - DLM overview
  - Bayesian Inference Networks in DLM
- Discussion of psychometric models that parallel DLM BINs
  - Comparison of terminology
- DLM psychometric model and estimator
- Initial results

THE DYNAMIC LEARNING MAPS PROJECT

DLM State Membership Map





Portion of Mathematics Learning Map

Key features of the DLM Project

- Instructionally-embedded assessments
- Instructionally relevant testlets
- Fine-grained learning maps
- A subset of particularly important nodes that serve as content standards – Essential Elements
- Accessibility and alternate pathways
- Dynamic assessment
- Status and growth reporting that is readily actionable
- Professional development
- A technology platform to tie it all together

THE UNIVERSITY OF KANSAS

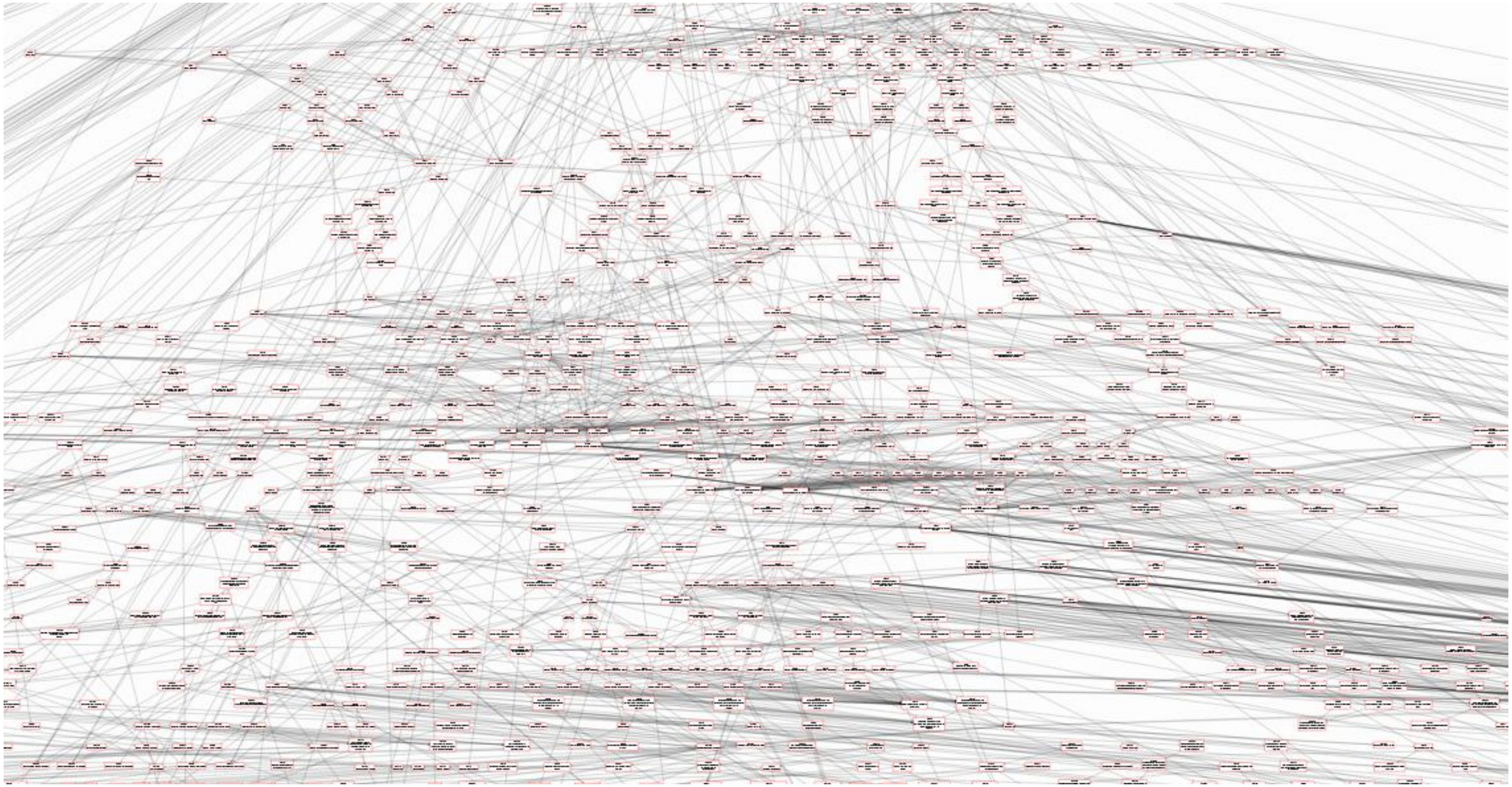
Instructionally Embedded Tests

- Assessment is most useful when it is designed to help teachers help students learn
  - Better to modify the assessment than modify the instruction
  - Potentially easier to monitor growth

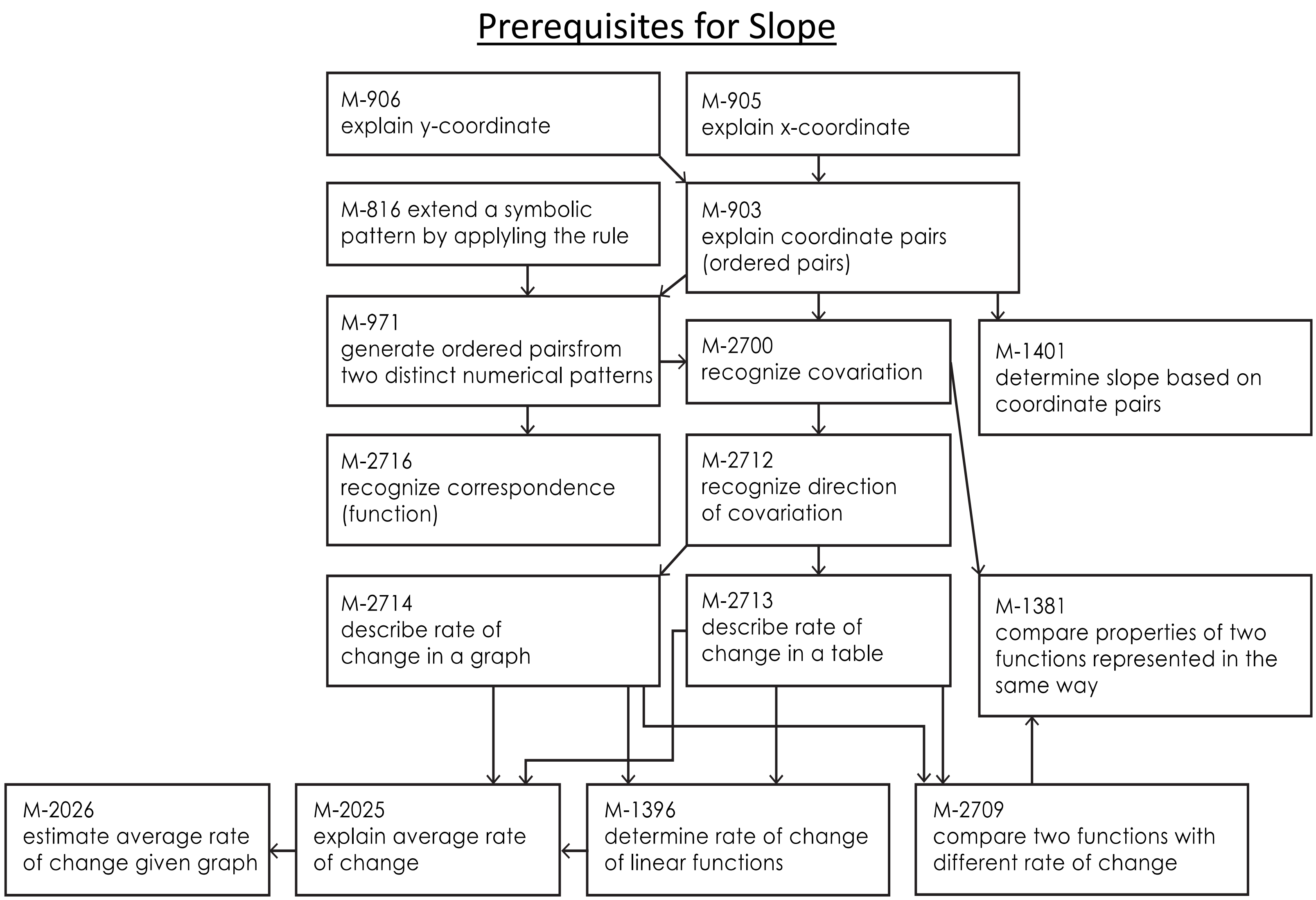
- Example: one task every other week for 30 weeks for a total of about 60 items
  - Compared to a typical summative alternate assessment with perhaps 30 items

THE LEARNING MAP

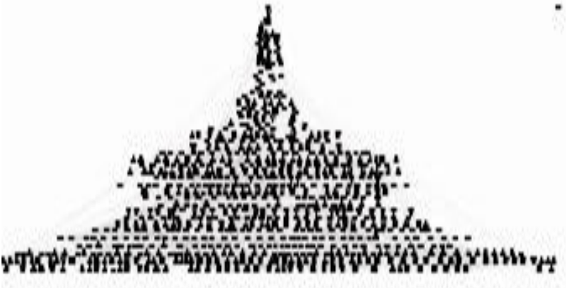
A Portion of the Math Map

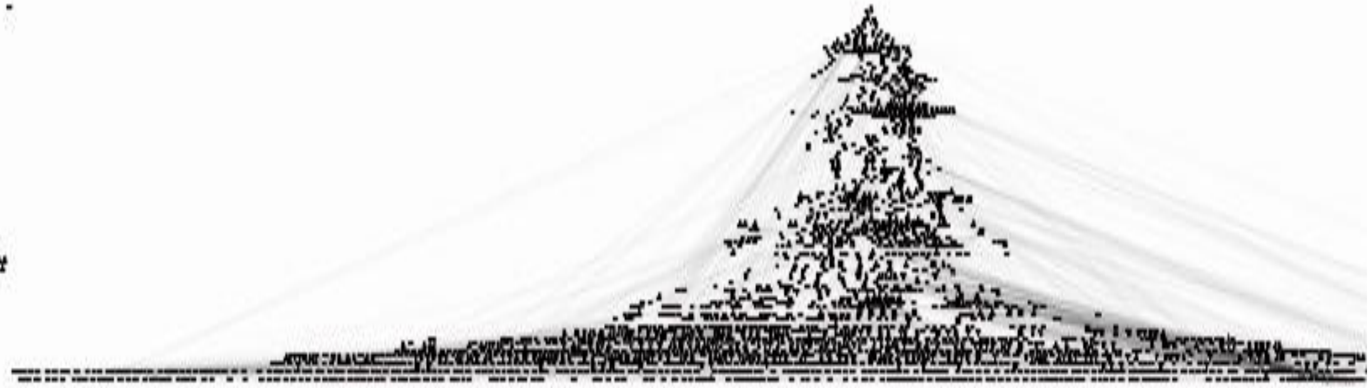


Zooming In on a Portion of the Math Map



Quick Facts about the Map





- English Language Arts
  - 141 foundational nodes
  - 1,645 ELA nodes
    - 538 Essential Elements
  - 3,982 edges/connections
- Mathematics
  - 141 foundational nodes
  - 2,312 mathematics nodes
    - 172 Essential Elements
  - 4,838 edges/connections

Common to both

Items developed and tested on only a small set of these nodes currently

DLM Terminology

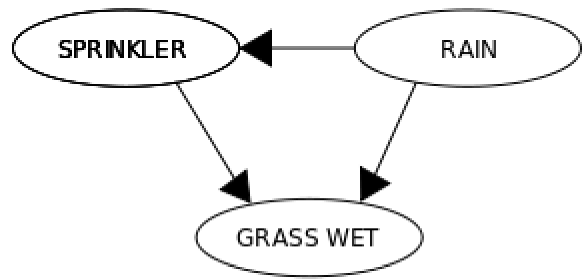
- DLM Terminology: straight from Bayesian Inference networks and graphical models
- Nodes: categorical latent variables
  - Analogs to latent factors in factor analysis or item response theory
- Nodes can be Parents or Children
  - Parents: Not predicted by anything (we would call this an Exogenous variable)
  - Children: Predicted by parents (we would call this an Endogenous variable)
- Edges: conditional dependencies between:
  - Nodes
  - Nodes and items
- The DLM "Learning Map" is called a Markov Blanket
  - Also called a Directed Acyclic Graph or DAG

Woefully Short Primer on Bayesian Networks

- BINs describe multivariate data using conditional probabilities

Image source: Wikipedia (yeah, that Wikipedia)

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

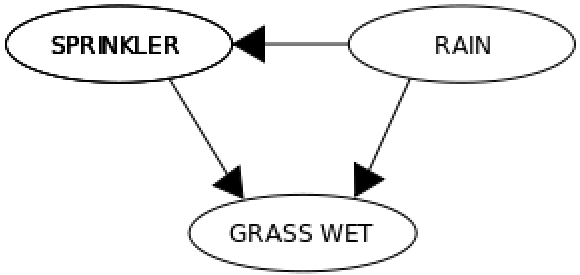
		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

- In the image, three variables observed:
  - Did it rain?
  - Were the sprinklers on?
  - Was the grass wet?

- The BIN includes the set of parameters leading to the probabilities in the tables

Woefully Short Primer on Bayesian Networks

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Joint distribution of Rain, Sprinkler, and Grass Wet given by:

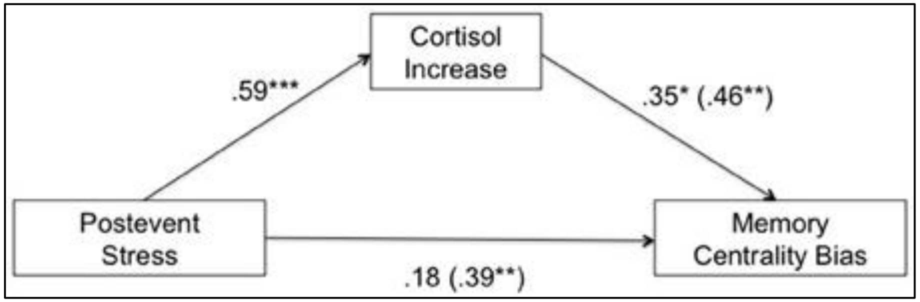
$$= P(Grass, Sprinkler, Rain)$$

$$P(GrassWet = T|Rain, Sprinkler)P(Sprinkler|Rain)P(Rain)$$

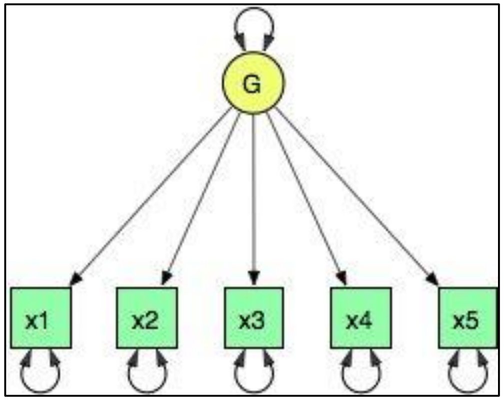
- Conditional/Marginal distribution of each variable: Bernoulli
- This example has all observed variables, but latent variables can also be defined
  - Hidden/unobserved nodes

Worlds Colliding….Psychometric Models are BINs

- Here are some BINs that may be more familiar in the social sciences…



Conditional/Marginal distribution of each variable: Normal
Nodes: Observed variables (or more specifically, X, Y, and M)



Conditional/Marginal distribution of each variable: Normal
Nodes: 5 Observed variables (X1 – X5)
1 unobserved variable (G)

More BIN Terminology

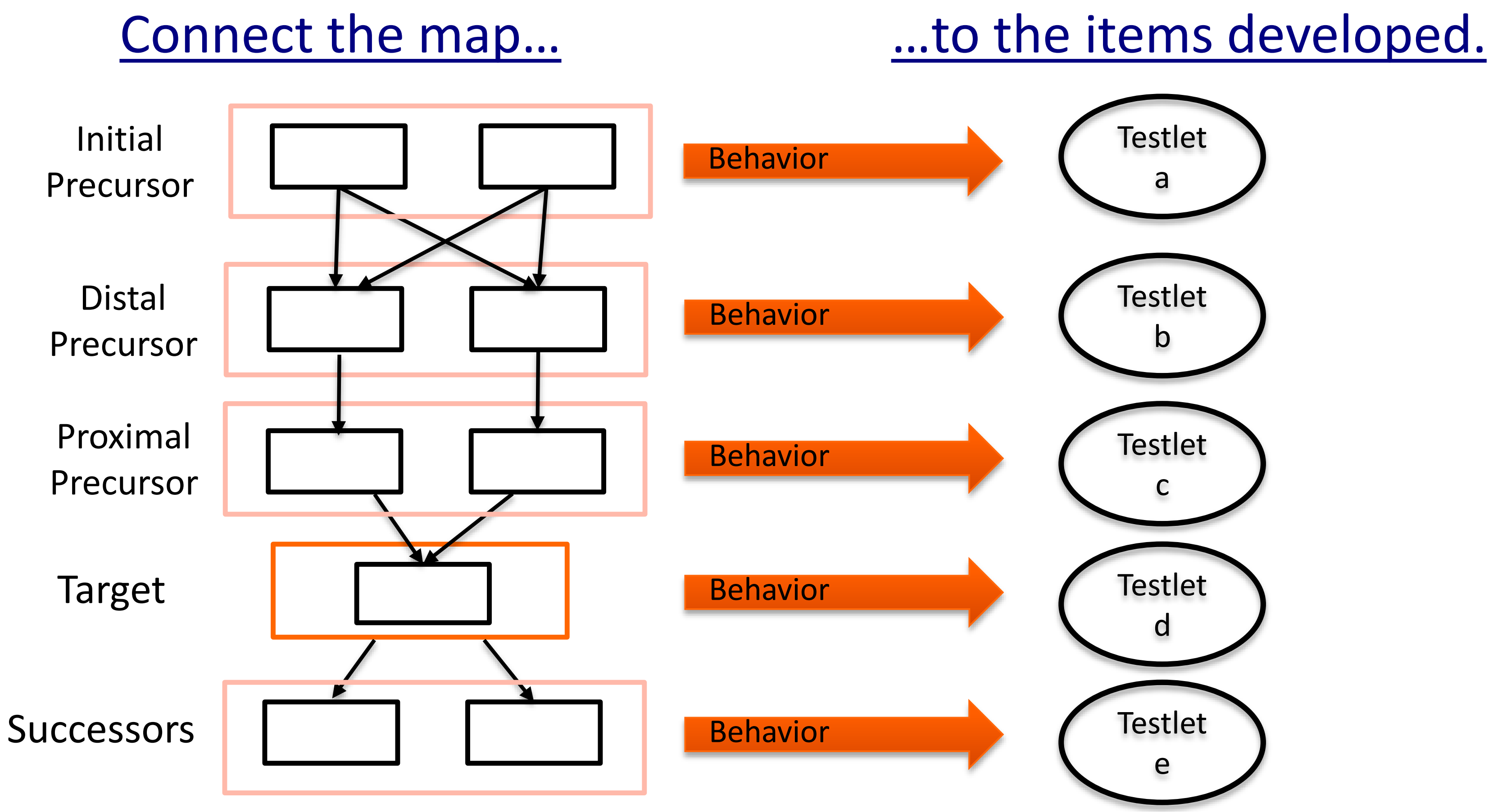
- Network Learning/Training = Estimation of model parameters
  - Often done with Bayesian/MCMC where priors are placed on nearly all parameters
- Estimation typically done using cross-validation
  - Estimation on one/several samples of data
  - Prediction done with left-out samples of data
- From Psychometrics: Model fit…not evaluated in same way
  - BIN model fit based on:
    - Prediction of left-out samples
    - Posterior predictive checks
    - Entropy (for categorical hidden nodes)
      - This is like saying your CFA model fits because your Omega reliability coefficient is high

ANOTHER DLM CATCH: ITEMS ARE IN TESTLETS

Item Types

- Single-select multiple choice
- Multi-select multiple choice
- Technology enhanced:
  - Sorting
  - Matching
  - Hot text (ELA)
- Teacher observation*
- Extended performance event*

Testlets in Linkage Levels



DLM FROM A PSYCHOMETRICS PERSPECTIVE

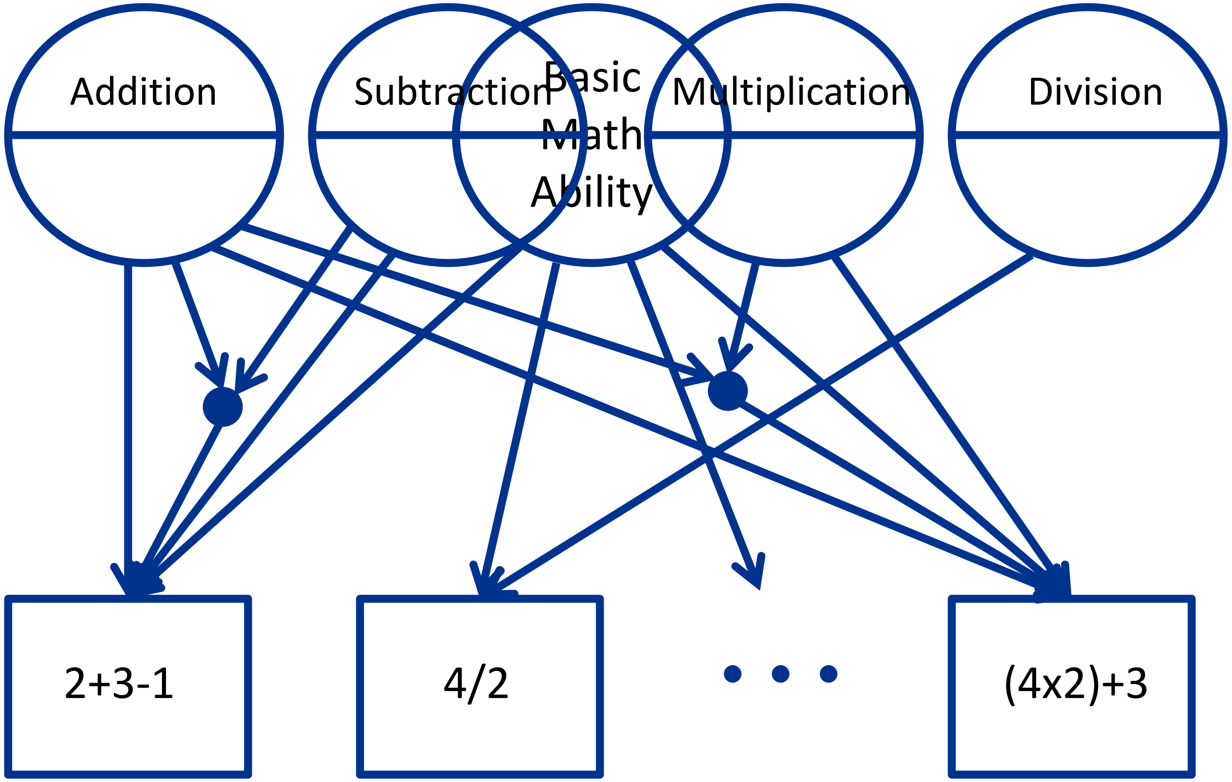
Overall Goal: Develop Method to Evaluate Students

- DLM Project features a learning map
- “Nodes” on map are student-specific
  - Informs instruction – what a student knows or does not know
- Process of determining student status is called “cognitive diagnosis”
- Term for larger set of psychometric tools that fall under an family of models that I call diagnostic classification models or DCMs

Diagnostic Classification Model Names

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

Path Diagram of Traditional Psychometrics vs. DCMs



THE UNIVERSITY OF KANSAS

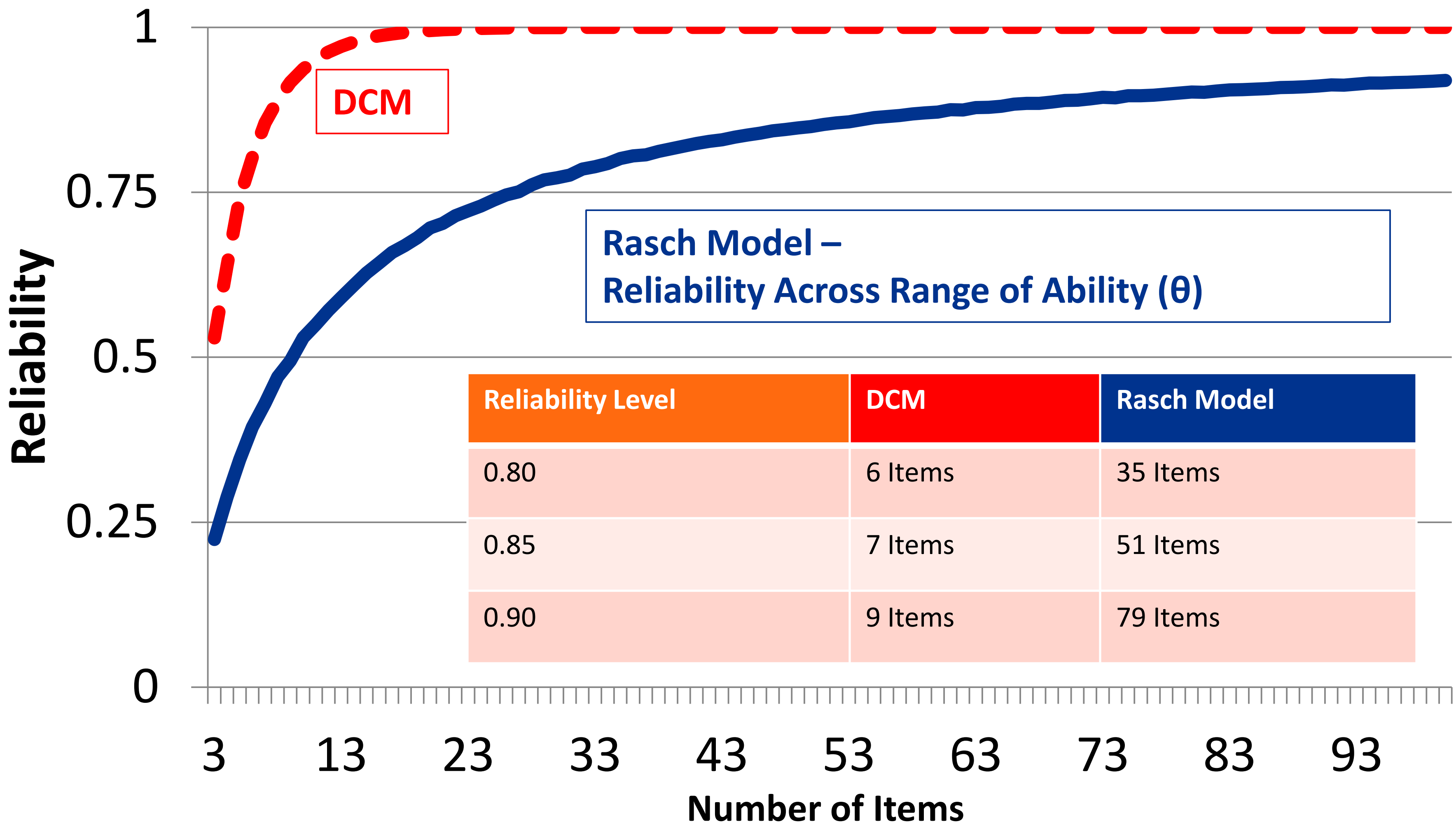
Multiple Dimensions of Ability

- The set of nodes in the DLM "learning maps" represent the multiple dimensions of ability

- Other psychometric approaches have been developed for multiple dimensions
  - Multidimensional item response models
  - Subscores in classical test theory

- So…why not use something more familiar in testing?
  - Reliability of estimates is often poor for practical test lengths
  - Dimensions are often very highly correlated
  - Large samples are needed

Example Theoretical Reliability Comparison



Reliability Level	DCM	Rasch Model
0.80	6 Items	35 Items
0.85	7 Items	51 Items
0.90	9 Items	79 Items

Templin, J. & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30,* 251-275.

DLM MODELING STRATEGY

Modeling Strategy: Text Description

- We consider items to be nested within testlets which interact with students

- The item model combines the loglinear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009) with a crossed random effect (e.g. Van den Noortgate, De Boeck, & Meulders, 2003) for within-testlet dependencies

- The functional form of the model resembles the LCDM/IRT combination model of Templin (in press) as used for testlets (Jurich and Bradshaw, under review)

DLM Measurement and Structural Models

Measurement Model (Items)

- Logit of a correct response ($X_{ei} = 1$) to item i by examinee e:

$$\text{logit}\big(P(X_{ei} = 1 | \alpha_{en})\big) = \lambda_{i,0} + \lambda_{i,1,(n)}\alpha_{en} + \gamma_{ei(t)}$$

- Where:

$$\gamma_{ei(t)} \sim N(0, \sigma^2_{\gamma_t})$$

Structural Model (Nodes/Map Edges)

- Marginal node distribution is:

$$\alpha_{en} \sim B\left(p_{\alpha_n} = P(\alpha_{en} = 1 | \alpha_{en'})\right)$$

- With model for p_{α_n} conditional on precursor nodes:

$$\text{logit}\big(P(\alpha_{en} = 1 | \alpha_{en'})\big) = \lambda_{n,0} + \lambda_{n,1,(n')}\alpha_{en'}$$

Model Parameter Descriptions

$$\text{logit}\left(P(X_{ei} = 1|\alpha_{en})\right) = \lambda_{i,0} + \lambda_{i,1,(n)}\alpha_{en} + \gamma_{ei(t)}$$

Here:

- α_{en} is the mastery status of examinee e on node n
  - For masters, $\alpha_{en} = 1$; for non-masters $\alpha_{en} = 0$
  - The value of α_{en} is arbitrary: masters are not 1 unit higher in ability
- $\lambda_{i,0}$ is the intercept
  - The log-odds of a correct response for non-masters when the testlet interaction is zero (average)
- $\lambda_{i,1,n}$ is the “main effect” of mastery of node n
  - The difference in log-odds of correct response between masters and non-masters of the node
- $\gamma_{ei(t)}$ is the testlet effect for examinee e and testlet t

Structural Model Parameters

$$\text{logit}\Big(P(\alpha_{en} = 1|\alpha_{en'})\Big) = \lambda_{n,0} + \lambda_{n,1,(n')}\alpha_{en'}$$

Here:

- $\alpha_{en'}$ is the mastery status of examinee e on node n'
  - For masters, $\alpha_{en'} = 1$; for non-masters $\alpha_{en'} = 0$
  - The value of $\alpha_{en'}$ is arbitrary: masters are not 1 unit higher in ability
- $\lambda_{n,0}$ is the intercept
  - The log-odds of the probability of mastery for node α_{en} for non-masters of node $\alpha_{en'}$
- $\lambda_{n,1,(n')}$ is the "main effect" of mastery of node n'
  - The difference in log-odds of mastery for node α_{en} between masters and non-masters of node $\alpha_{en'}$

THE UNIVERSITY OF KANSAS

The DLM Model Estimator

- Estimation via Metropolis-Hastings algorithm
  - Dates to physical chemistry (1950)

- All item/map parameters use uniform prior distribution

- All student node parameters had a prior distribution described by map parameters
  - We would call this an empirical prior in the Bayesian world
  - This is what is used in standard CFA/MLM/MIRT, etc…

- All testlet effects had a prior of a normal distribution with zero mean and estimated testlet variance

THE UNIVERSITY OF KANSAS

Wrapping Up

- The DLM project is an ambitious attempt to measure a lot of things simultaneously
- Early results show that more work is needed to ensure stable estimates of parameters are present
- Some of the map results suggest some nodes/attributes/factors aren't present