Mixtures of Diagnostic Skill Profile Models

Matthias von Davier*

@

CILVR Conference on Mixture Models, May 18-19,

University of Maryland, College Park

*Educational Testing Service, Center for Statistical Theory and Practice

The presentation uses joint work with:

- Henry Braun
- Alina von Davier
- Xiaomin Huang
- Xueli Xu
- Kentaro Yamamoto

Overview

- What are Diagnostic Models, and why extend them?
- The General Diagnostic Model (GDM)
- Multiple Population and Mixture GDMs
- Scale linkage across GDMs
- Applications

What are Diagnostic Models?

- Models for reporting skill profiles
- Multiple skills, discrete levels, often mastery/non-mastery
- Models are often specified for dichotomous items
- Design matrix (Q-matrix) relates skills to items

DM are LCA, MIRT, DINA, NIDA et al.:

- Constrained latent class models
- Discrete M-IRT, latent response models
- DINA, Deterministic Input, Noisy AND (OR etc.)
- NIDA, Noisy Input, Deterministic AND (OR etc.)
- NOW: General Diagnostic Model, or maybe:
- Multidimensional Discrete Latent Trait Models (mdltm)

Mixture Diagnostic Models are useful:

1. For scale linkage across test forms and populations
2. For studying DIF using multiple populations
3. For examining appropriateness of Q-matrix definition
4. As “poor-researchers” conditioning model

von Davier & Yamamoto (2004) develop a general diagnostic model (GDM) framework. The GDM uses ideas from M-IRT and Multiple-Classification & Located-Latent -Class-Models:

- Allows polytomous items, dichotomous items, mixed in a form

- Allows polytomous, mastery/non-mastery, pseudo-continuous skills

- von Davier (2005) describes partial credit GDM, develops EM algorithm

- 2006: Extension to mixture and multiple group GDMs

The partial credit version of the GDM is:

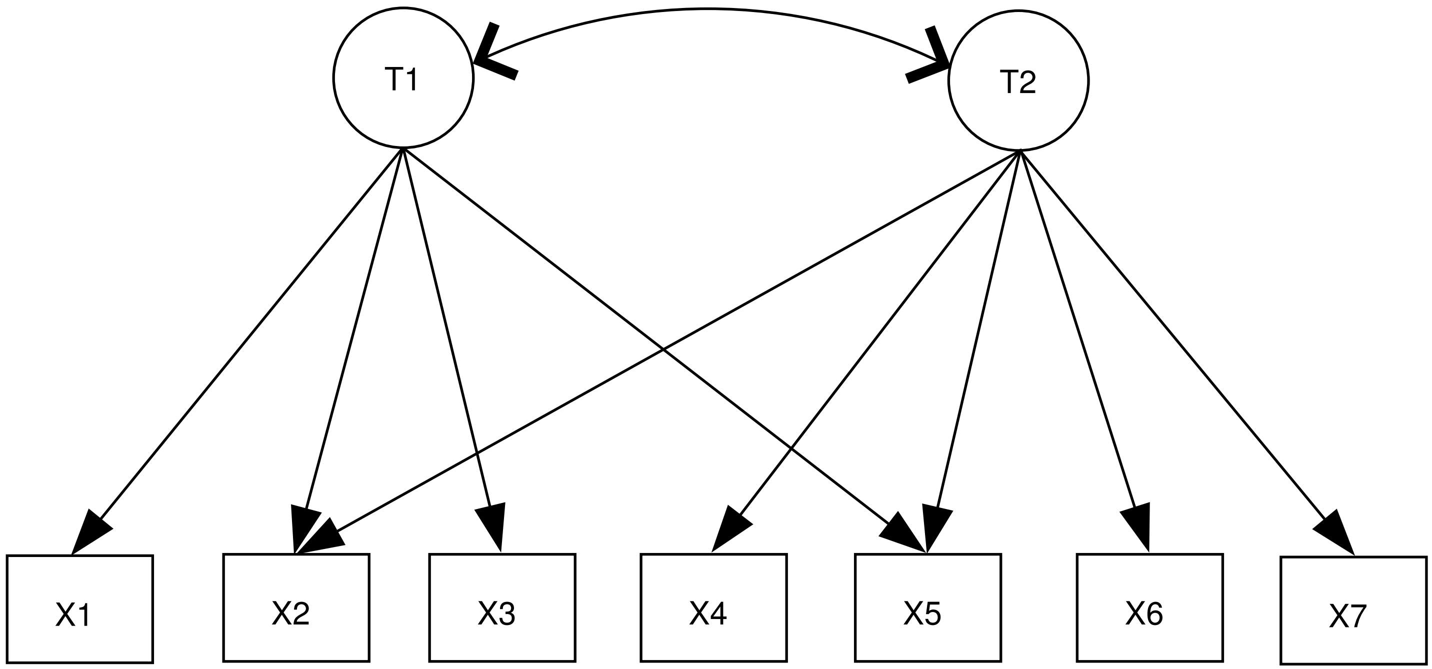
$$P(X = x \mid \beta_i, a, q_i, \gamma_i) = \frac{\text{exp}\left[\beta_{xi} + \sum_{k=1}^{K} x\gamma_{ik} q_{ik} \theta(a_k)\right]}{1 + \sum_{y=1}^{m_i} \text{exp}\left[\beta_{yi} + \sum_{k=1}^{K} y\gamma_{ik} q_{ik} \theta(a_k)\right]}.$$

with item difficulties β_i, slopes γ_{ik}, skills a_k, levels θ_k, Q-matrix $(q_{ik})_{i,k}$ for $i = 1 \dots I$ and $k = 1 \dots K$.

A (rather small) diagnostic model example:

- Two skills, e.g. dichotomous $T1 \in \{-1, 1\}$ and ordinal $T2 \in \{-2, -1, 0, 1, 2\}$
- Seven items, a mix of dichotomous $X1..X3 \in \{0, 1\}$ and polytomous $X4..X7 \in \{0, 1, 2, 3\}$
- Q-matrix $((1110100)^T, (0101111)^T)$

An illustration of the above example:



Single Population Model

Without mixtures / multiple populations, we assume:

- Parameters of the diagnostic model hold for all examinees, i.e., the same difficulty and slope parameters can be used for everyone

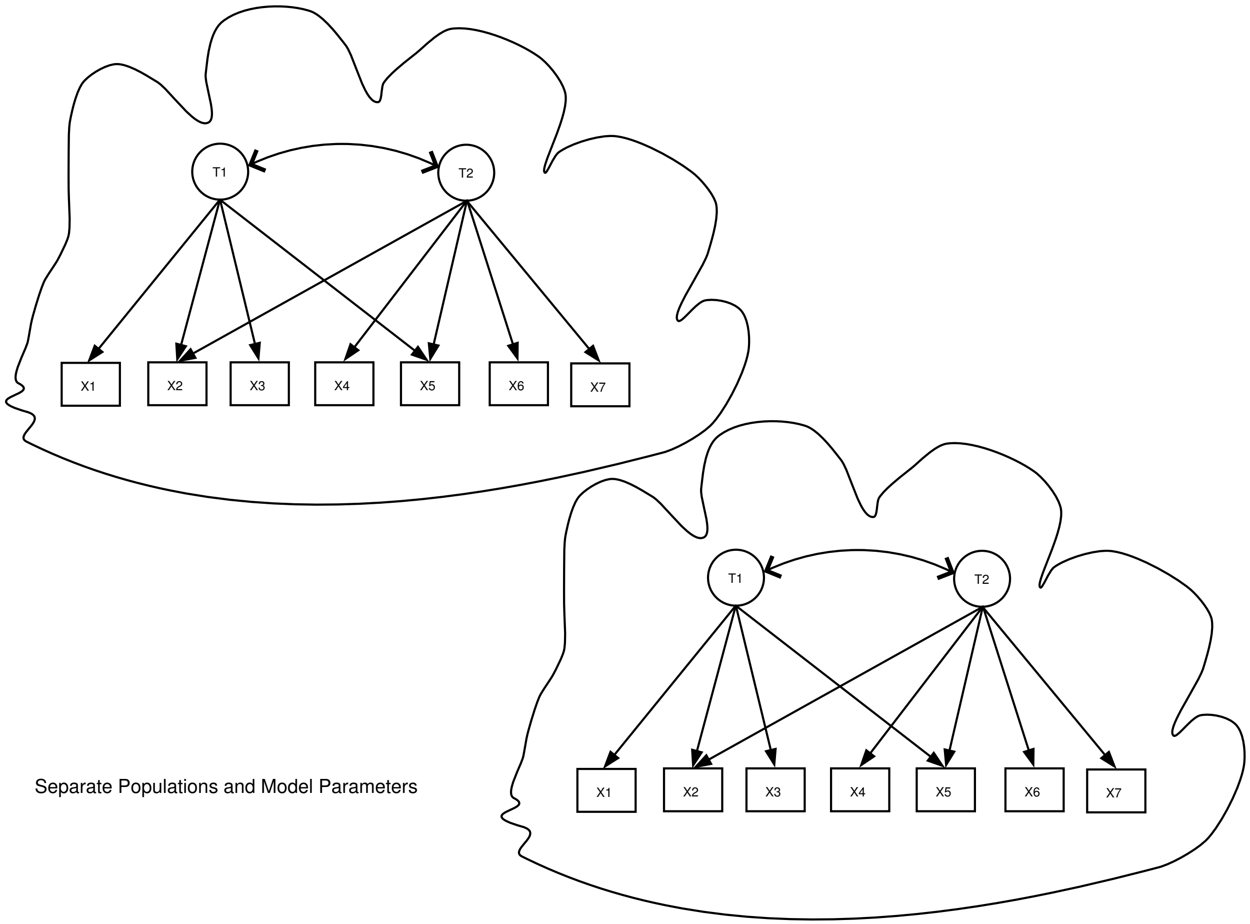
- A single examinee ability distribution (there are no covariates of ability), that is, knowledge about other variables is either unavailable or is assumed irrelevant.

The mixture / multiple-group version of the GDM:

$$P(X = x \mid \beta_i, a, q_i, \gamma_i, g) = \frac{\text{exp}\left[\beta_{xig} + \sum_{k=1}^{K} x\gamma_{ikg} q_{ik} \theta(a_k)\right]}{1 + \sum_{y=1}^{m_i} \text{exp}\left[\beta_{yig} + \sum_{k=1}^{K} y\gamma_{ikg} q_{ik} \theta(a_k)\right]}.$$

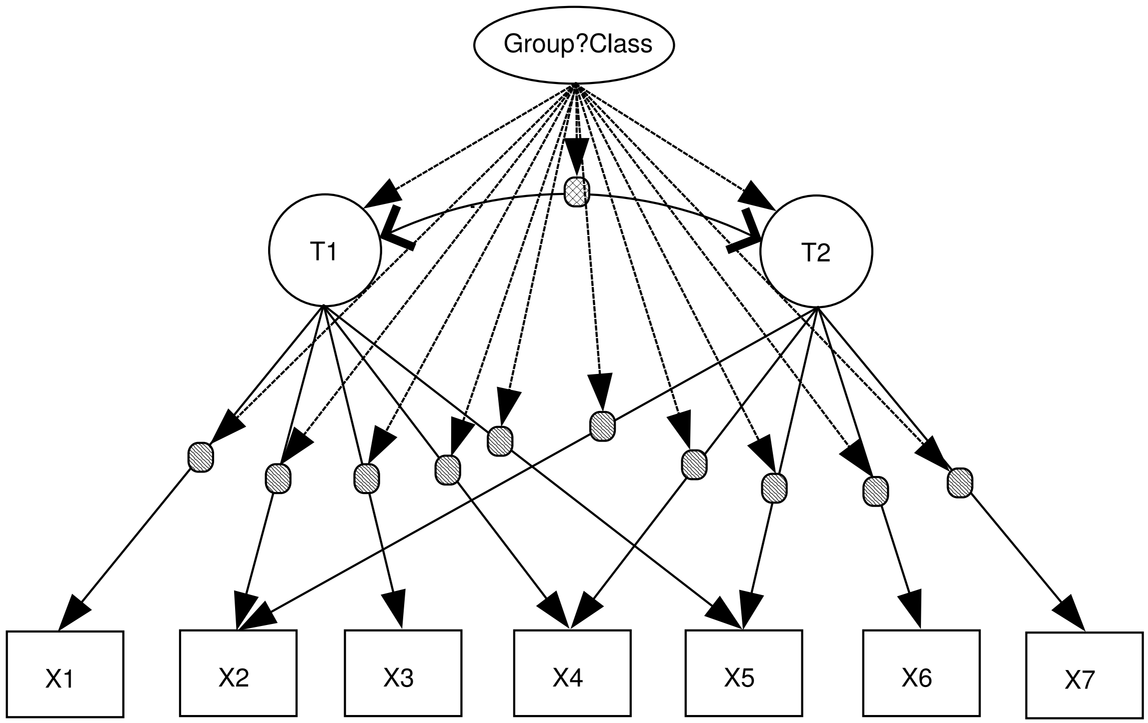
with parameters as defined above, and added group index g.

Separate model parameters in separate groups:



Separate Populations and Model Parameters

Group indicator for separate model parameters:



Multigroup Model with Group Specific Item Parameters

What is a concurrent calibration model good for?

- Study how different populations are
- Unmix populations when different strategies or response styles are involved
- Identify 'unscalables', speededness etc.

Group indicator g separates model parameters:

- Group g is an observed variable in classical multiple-group models

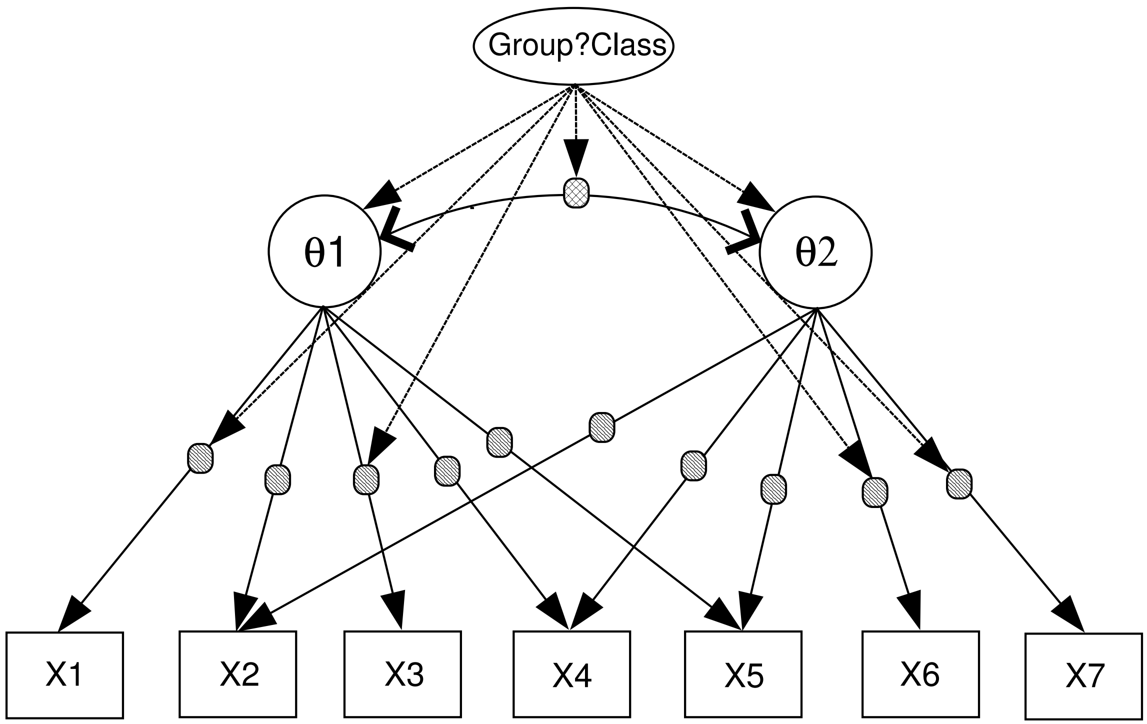
- Group membership can be unobserved -> mixture IRT (Yamamoto, '89; Mislevy & Verhelst, '90; Rost, '90; von Davier & Rost, '95; ...), latent Class models (Lazarsfeld & Henry 1968, Haberman ...)

- Classification into groups may be missing or unreliable -> partially missing grouping information (von Davier & Yamamoto, 2004)

Scale linkages across mixture / multiple group diagnostic models:

- Arrows originating from group indicator mean "depends on"
- Missing arrows mean "is independent of g, i.e., the same for all groups"
- *von Davier*[2] describe IRT scale linkages across groups as constrained maximization problem
- Can be applied here: Mixture / multiple group GDM's share a lot with constrained multiple-group IRT

A mixture / multiple group model with equality constraints:



General Diagnostic Model with Group Specific and Unspecific Item Parameters

Constraints across mixture components / multiple groups:

- Note: Equality constraints across all groups show up as non-arrows

- Actual implementation is the other way around: Specify what is equal!

- Parameter fixations and equality constraints allow complex linkages across groups (more complex than easily represented in graphs)

- For the GDM, these constraints allow the same or even different Q-matrices in different populations

Different Q-matrices in different populations:

1. Define a “super”-Q-matrix with “1” entries if a skill is needed for an item in at least one group, “0” otherwise

2. Impose slope parameter fixations (=0.0) for skills that are not needed in certain groups for certain items

3. Impose additional constraints and fixations as neccessary, or hypothesized

4. Compare fit of models with constraints with the unconstrained model (or the less constrained)

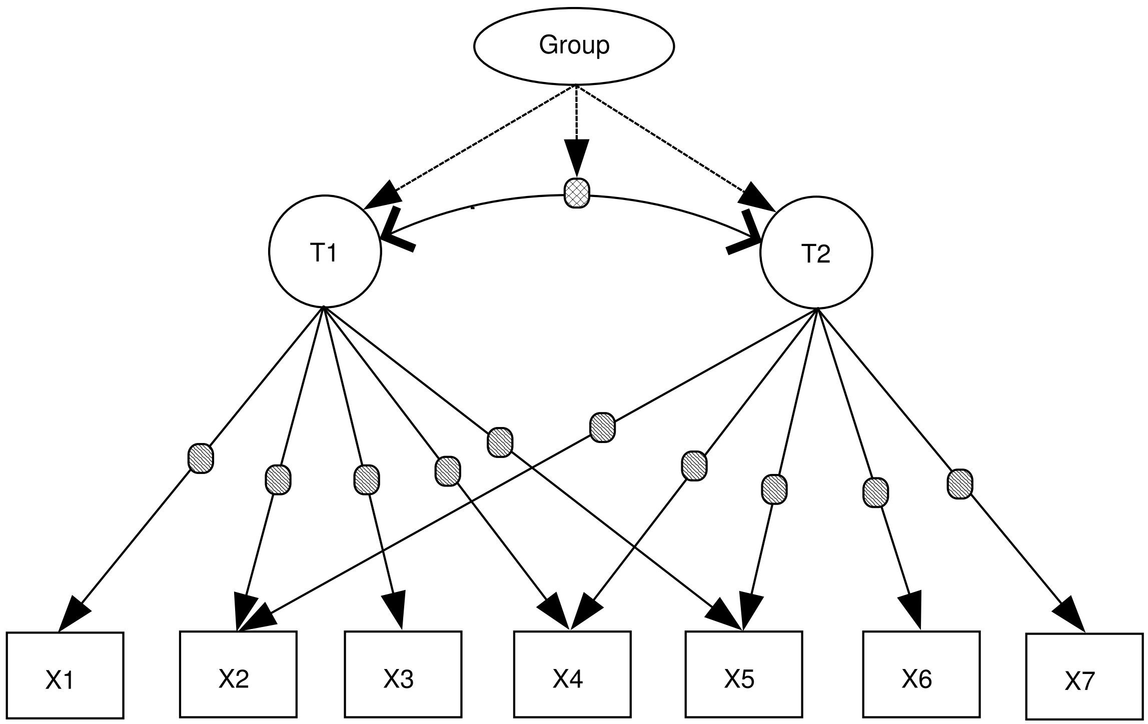
Why constrained mixture / multiple group models:

- For linking multiple forms (one anchor, multiple cohorts)
- Link chains of test forms (adjacent, but different anchors)
- Find subsets of grouping variable with similar constraints
- Study differences when multiple languages are involved

Strongest form of linkage across multiple populations:

- One set of item parameters, the same across all groups
- Only ability distributions [here $P(T1, T2|g)$] differ across groups
- This model measures identical skills allowing different skill distributions across groups
- See applications section below...

Strongest form of linkage across multiple populations:



Multigroup Model with Group Unspecific Item Parameters

Why models with same item parameters across groups:

- Link different administrations with the same items
- Assess differences in ability distributions across groups
- Use as “poor-researchers” conditioning model
- Baseline model. Start here, relax constraints if necessary

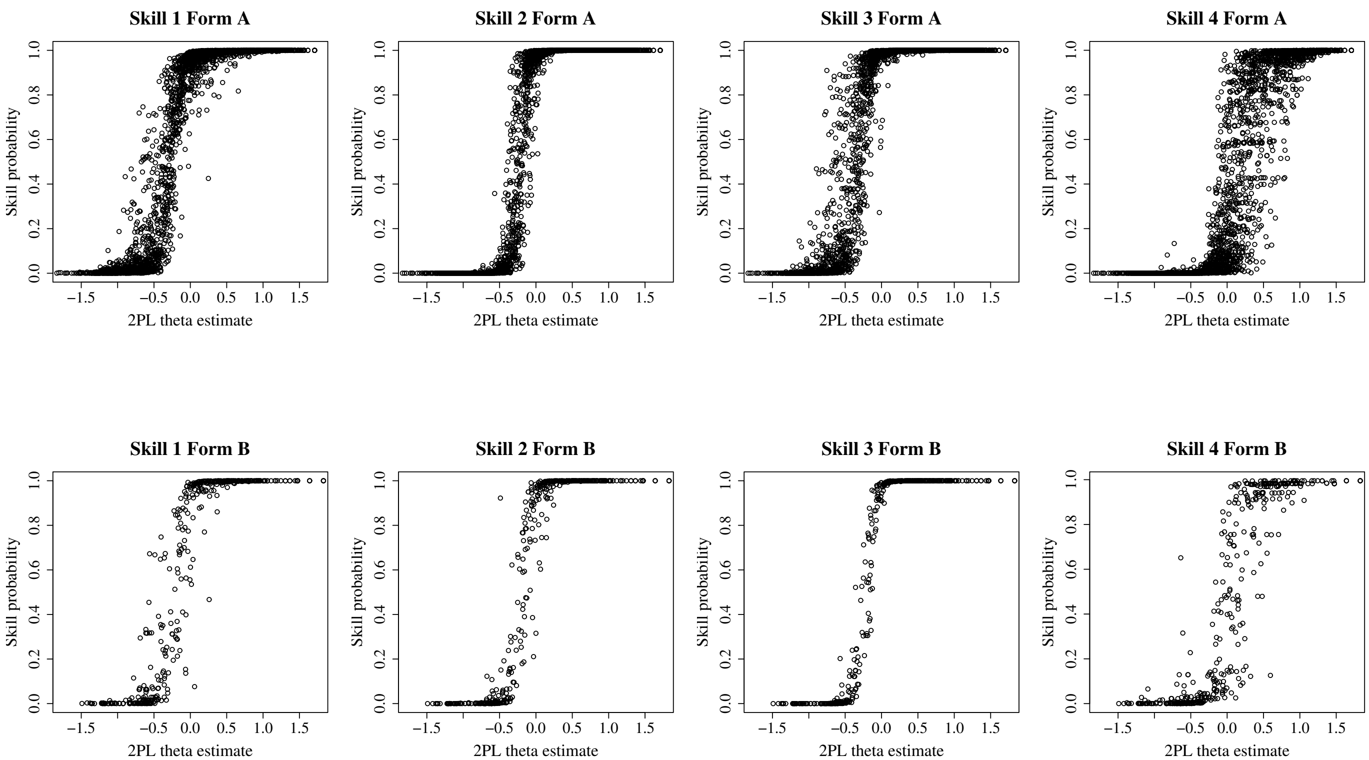
Applications of General Diagnostic Models (GDMs)

- English Language Testing
- National Large Scale Assessment
- International Assessments
- K-12 Accountability Testing

GDMs and English Language Testing (von Davier; 05)

- Uses TOEFL iBT pilot data
- Compares GDM and 2PL/GPCM
- 1-dim. IRT model fits as good as GDM
- Parsimony (Occam's Razor) favors 1-dim. IRT
- 2-dim. IRT fits Reading & Listening joint data

English Language GDM, Listening Form A & B:



Xu & von Davier (2006) use a multiple-group GDM for Large Scale Survey Data. One may use gender, race and other variables as a grouping variable.

- Data from 2002 12th grade NAEP assessments
- Reading (3 dimensions), Math (4 + 3 dimensions)
- Data extremely sparse; complex student & item sample
- Parameter recovery study supports results

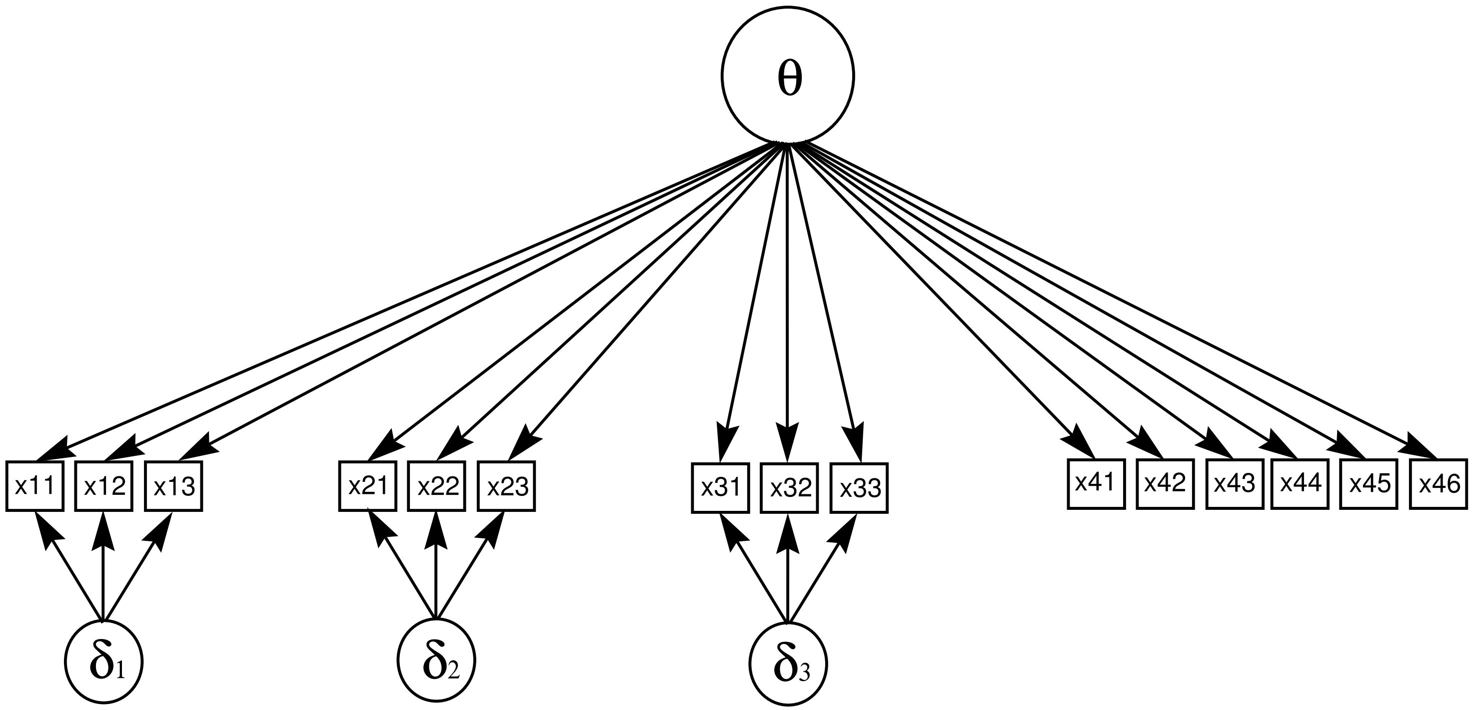
Xu & von Davier (2006) study parameter recovery of the GDM under different levels of sparseness:

Missing	Measure	10%	25%	50%
Item Parameter	Average Bias	0.001	0.002	0.005
	Average RMSE	0.071	0.083	0.119
Skill Distribution	Average Bias	0.000	0.000	0.000
	Average RMSE	0.004	0.004	0.007

Huang & von Davier (2006) use mixture IRT, GDMs, and Latent Class Models:

- Data based on ~47,000 adults from 7 countries
- Background data from a survey on adult literacy
- Goal: Develop indicator variables using LCA, GDM and IRT
- Purpose driven model selection becomes crucial:
- LCA, IRT and GDMs fit short scales (almost) equally well

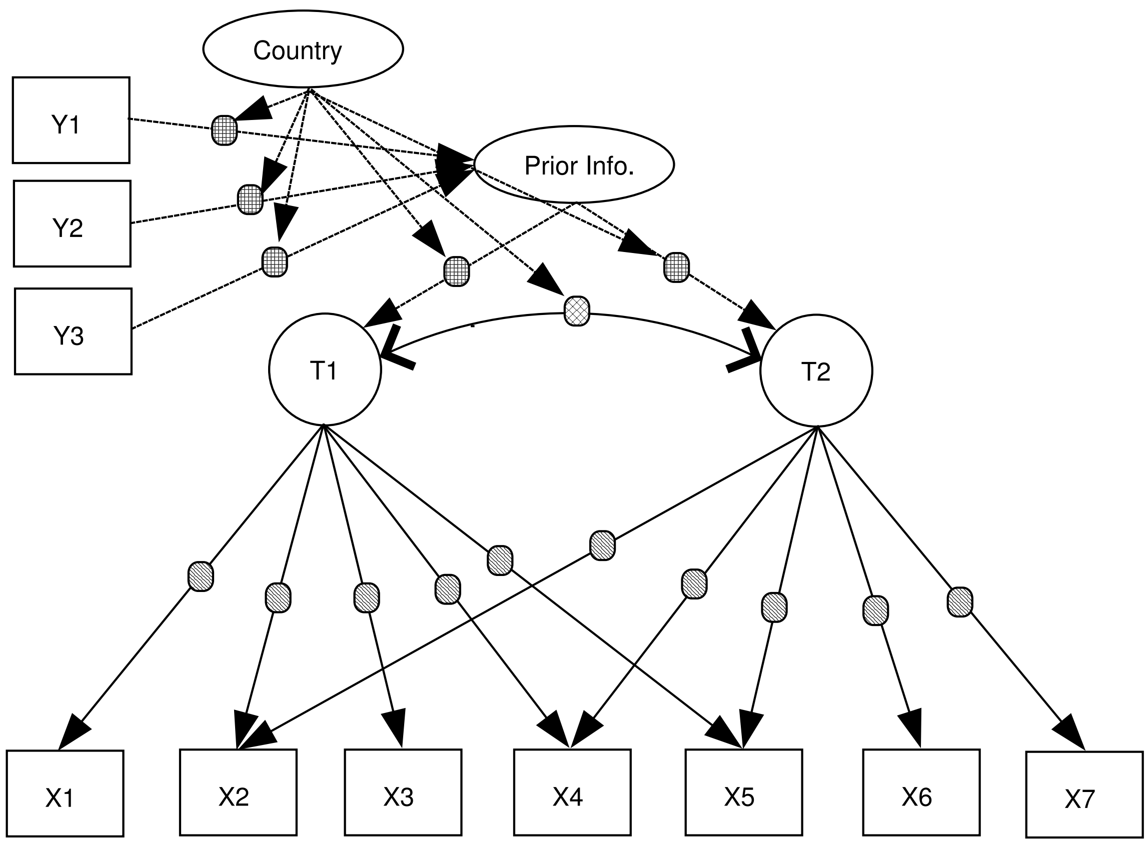
Braun & von Davier (forthcoming) use GDMs in K-12 arena:



Next steps:

- Include covariates for predicting skill distributions
- Use latent regression - conditioning in NAEP language
- Compare latent regression to multiple–group approach
- Develop parametric skill distribution models
- Research on model-data-fit & parsimony

Next steps in a picture:



Next? Model with Different Latent Regression in Different Countries

Summary: Mixture GDMs can be used to model:

- Single population general diagnostic models (GDM's, incl. IRT and LCA)

- Simultanous calibration-GDM's, mixture GDM's

- Constrained mixture GDM's, using complex linkages

- GDM's with missing data in item and in grouping information

- Multiple-group GDM's, with all items linked across groups