A comparison of the predictive performance of continuous and class-based latent trait models

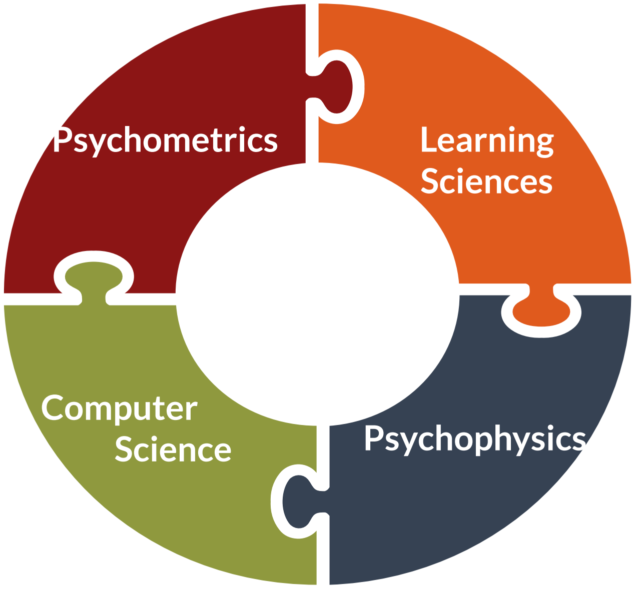
Anya Ma
wanjingm@stanford.edu
Stanford GRADUATE SCHOOL OF EDUCATION

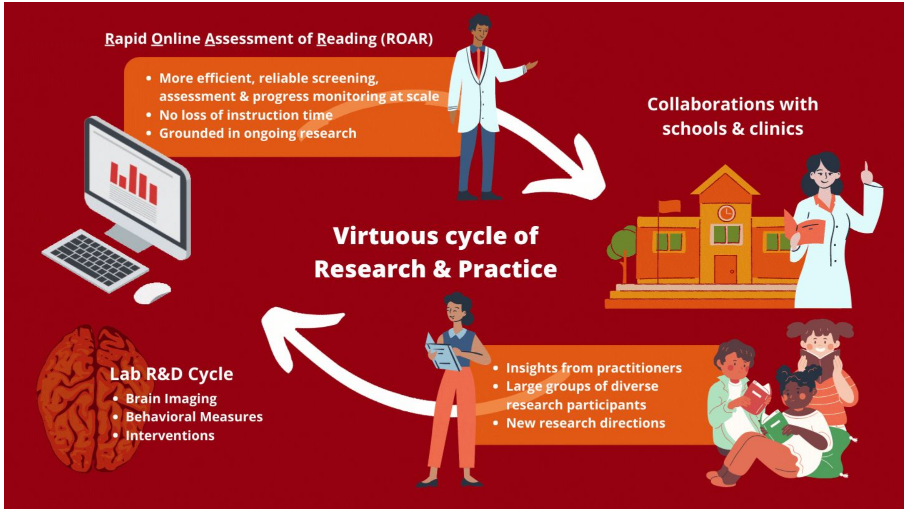
10/07/2025
BEAR Center Seminar, UC Berkeley

SLIDES

About My Research

- Former classroom teacher
- Interdisciplinary research for children with learning differences.





Rapid Online Assessment of Reading (ROAR)

- More efficient, reliable screening, assessment & progress monitoring at scale
- No loss of instruction time
- Grounded in ongoing research

Collaborations with schools & clinics

Virtuous cycle of Research & Practice

Lab R&D Cycle

- Brain Imaging
- Behavioral Measures
- Interventions

- Insights from practitioners
- Large groups of diverse research participants
- New research directions

The ROAR Assessment Suite

The ROAR is a tool for schools, clinics, and researchers. ROAR subtests under active research and development include:

Single Word Recognition

ROAR-SWR

ROAR-SWR measures a student's ability to **quickly recognize words**. Word recognition is at the foundation of reading ability and is important for reading fluency and comprehension.

Phonological Awareness

ROAR-PA

ROAR-PA measures **elision and sound matching** to assess a student's phonological awareness. This subtest is under active development and validation – please give it a try so that we can use your response to continue improving this measure.

Sentence Reading Efficiency

ROAR-SRE

ROAR-SRE measures students' ability to **silently read and understand sentences quickly and accurately**. This subtest is under active development and validation – please give it a try so that we can use your response to continue improving this measure.

Vocabulary

ROAR-Vocab

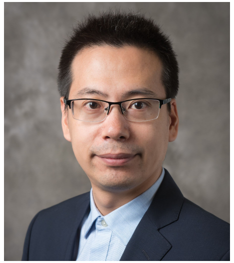
ROAR-Vocab measures **receptive vocabulary**, or the words that a student can recognize and correctly match to an image. This subset is under active development and validation – please give it a try so that we can use your response to continue improving this measure.

A comparison of the predictive performance of continuous and class-based latent trait models

PREPRINT

Wanjing Anya Ma[1], Yiqing Liu[1], Klint Kanopka[2], Wenchao Ma[3], Ben Domingue[1]

[1]Stanford University
[2]New York University
[3]University of Minnesota



Agenda

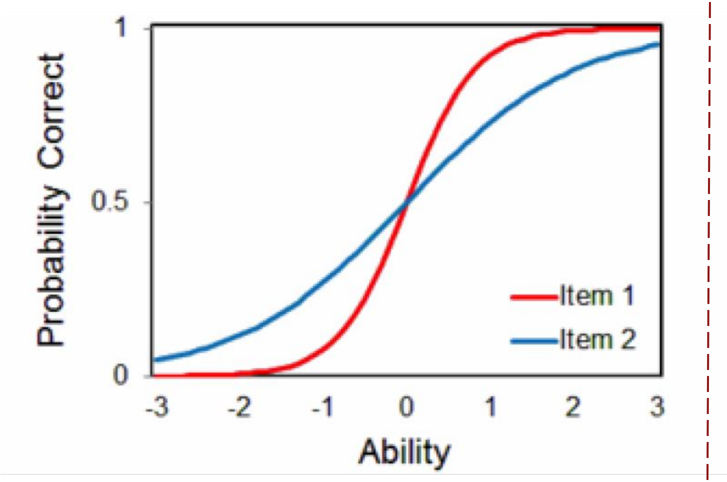
- **Motivation: rethinking about comparing IRT vs. CDM**
- Key Ideas: focus on out-of-sample predictive accuracy with empirical data
- CDM Basics
- Simulations: various conditions when data is generated by CDMs
- Empirical Datasets (N = 9)
- Discussion and future directions
- Q&A

Conceptions of ability

- The ability of a student can be conceptualized as:

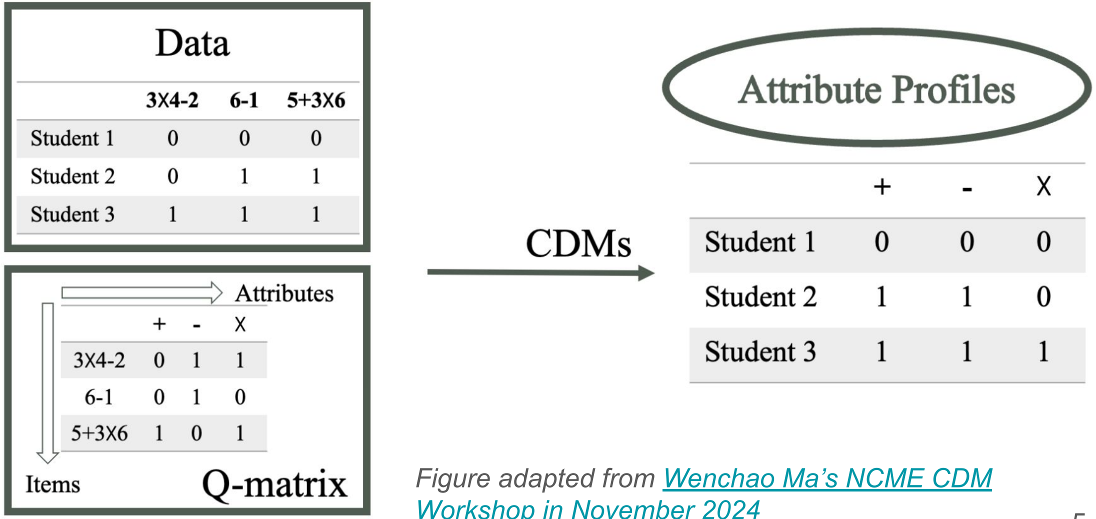
A **continuously** varying entity

(e.g., IRT models)



A bundle of **latent classes**

(e.g., Cognitive Diagnostic Models; CDMs)



Data

	3X4-2	6-1	5+3X6
Student 1	0	0	0
Student 2	0	1	1
Student 3	1	1	1

Q-matrix

Items	+	-	X
3X4-2	0	1	1
6-1	0	1	0
5+3X6	1	0	1

Attribute Profiles

	+	-	X
Student 1	0	0	0
Student 2	1	1	0
Student 3	1	1	1

Figure adapted from Wenchao Ma's NCME CDM Workshop in November 2024

Motivation I

- CDMs are appealing: potential to provide diagnostic inferences that can inform learning and teaching.
- They are also more “complex” than IRT!
- **Which model should we use — IRT vs. CDM?**
- Previous research in empirical datasets shows:

  CDMs fit better (Yamaguchi and Okada, 2018; Ma et al., 2020)

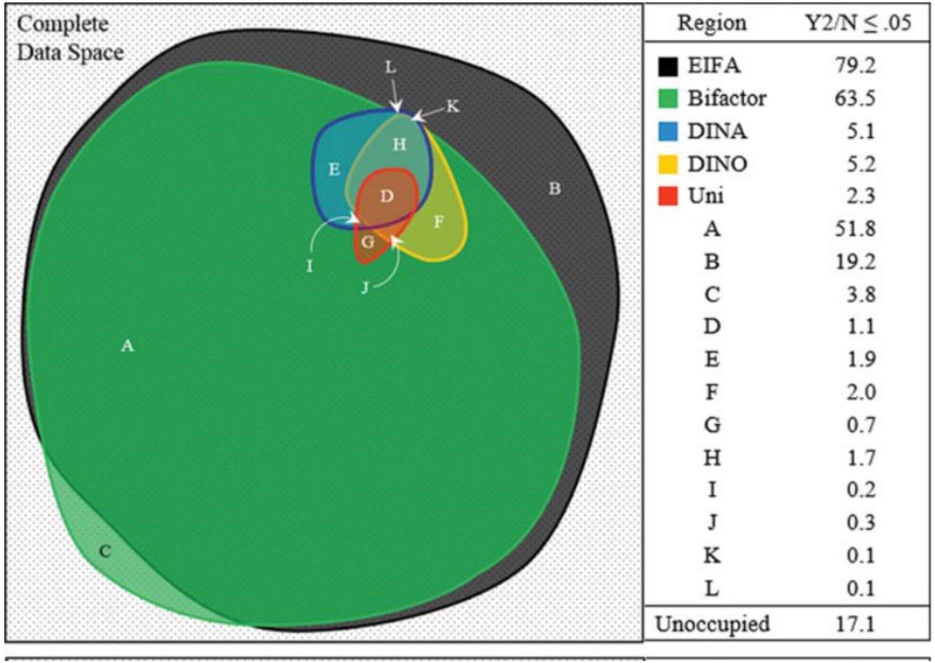
  CDMs fit worse: retrofitting issue (Templin & Bradshaw, 2014; von Davier, 2014)

- Challenges: they used (1) different datasets (2) with **in-sample** goodness of fit.
- Need a better way to examine this issue!

Motivation II: Model Complexity vs. Fitting Propensity

- Roberts & Pashler (2000): “models should not be judged only by how well they fit a data set; there also must be assessment of, and penalty for, flexibility” (p. 362)”

- Bonifay & Cai (2017): When the DGM is totally out of the context, DINA and DINO have higher fitting propensity then the 3PL.



Region	Y2/N ≤ .05
⬛ EIFA	79.2
🟩 Bifactor	63.5
🟦 DINA	5.1
🟨 DINO	5.2
🟥 Uni	2.3
A	51.8
B	19.2
C	3.8
D	1.1
E	1.9
F	2.0
G	0.7
H	1.7
I	0.2
J	0.3
K	0.1
L	0.1
Unoccupied	17.1

Figure reprinted from Bonifay & Cai, 2017

Agenda

- Motivation: rethinking about comparing IRT vs. CDM
- **Key Ideas: focus on out-of-sample predictive accuracy with empirical data**
- CDM Basics
- Simulations: various conditions when data is generated by CDMs
- Empirical Datasets (N = 9)
- Discussion and future directions
- Q&A

Key Idea I: focus on predictive accuracy

- Focus on **out-of-sample** predictive accuracy (Yarkoni & Westfall, 2017).

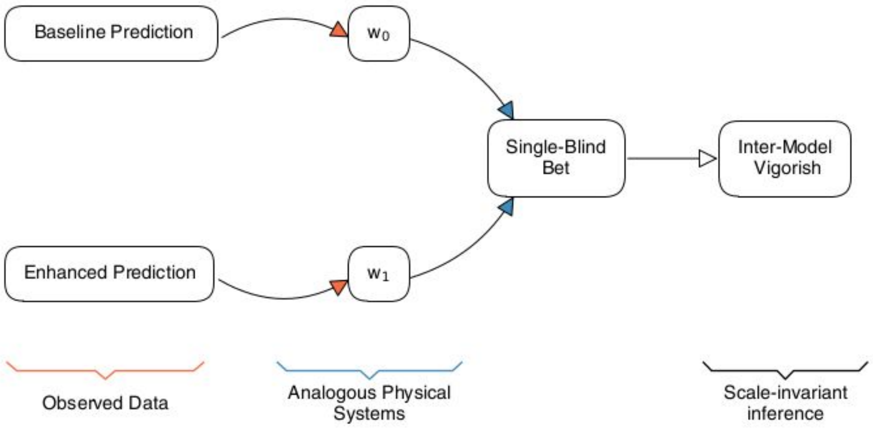
- The ***InterModel Vigorish*** (IMV; Domingue et al., 2024, 2025) to quantify the accuracy based on the improvement across two sets of predictions.

- Intuition: a dollar bet you expect to make a penny based on 'side information'

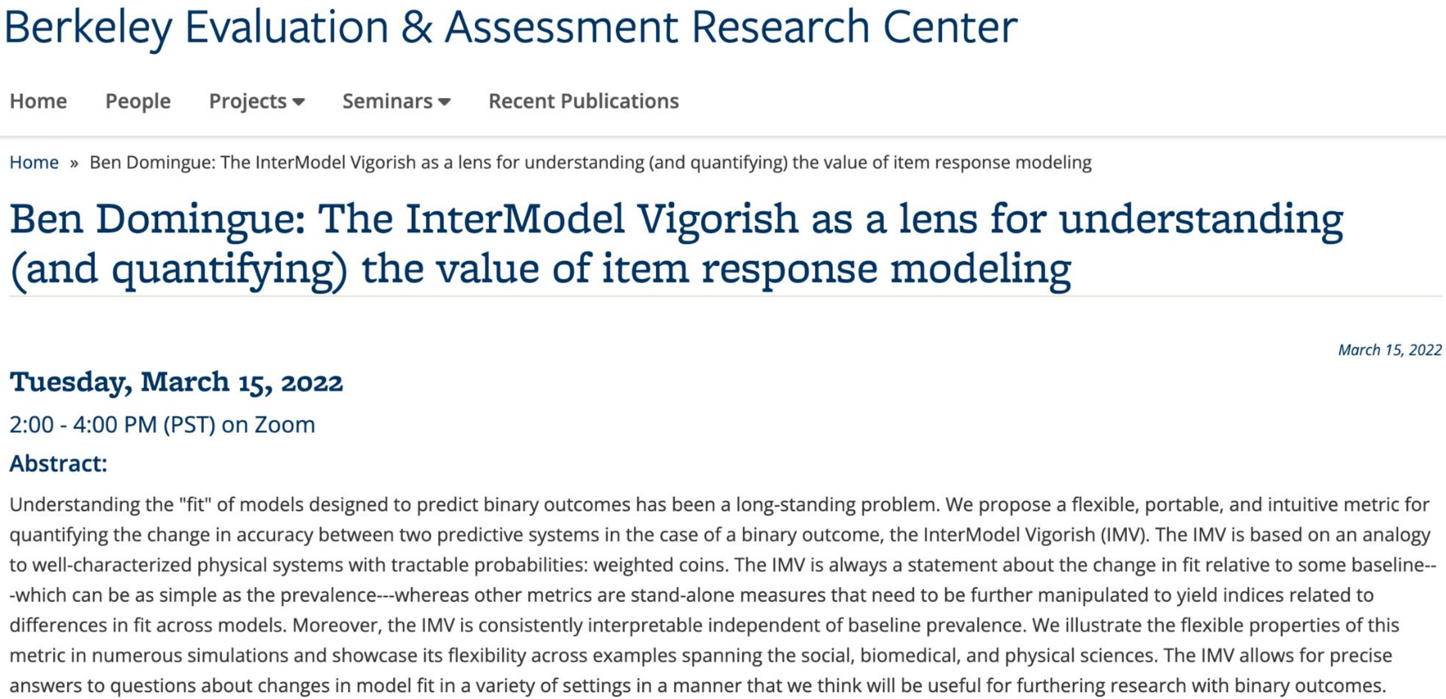
- IMV(m0,m1) = 0.1 → m1 outperforms m0

- Benchmark: IMV(1PL, 2PL) = 0.01

- Pros: used for the comparison of different models, portable, and generalizable



Learn more about the IMV?



Berkeley Evaluation & Assessment Research Center

Home » Ben Domingue: The InterModel Vigorish as a lens for understanding (and quantifying) the value of item response modeling

Ben Domingue: The InterModel Vigorish as a lens for understanding (and quantifying) the value of item response modeling

March 15, 2022

Tuesday, March 15, 2022

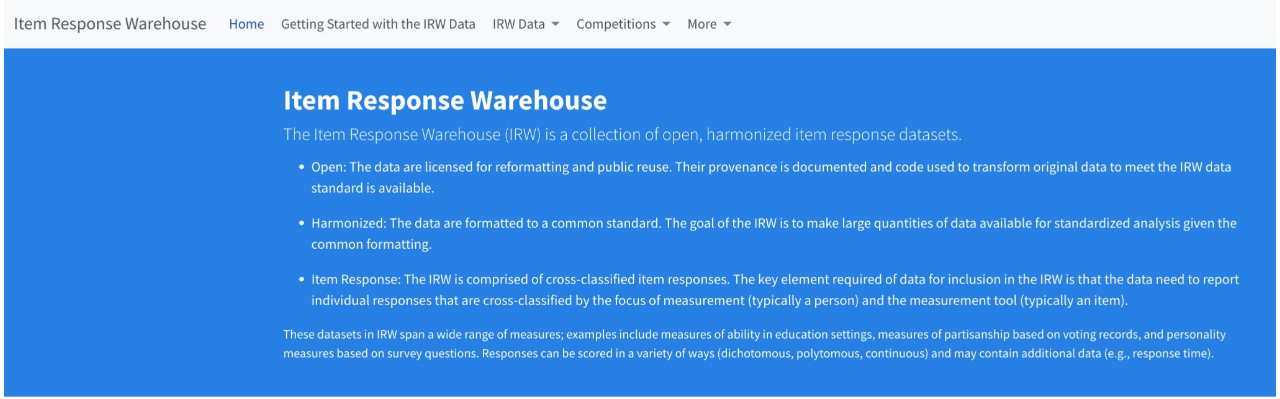
2:00 - 4:00 PM (PST) on Zoom

Abstract:

Understanding the "fit" of models designed to predict binary outcomes has been a long-standing problem. We propose a flexible, portable, and intuitive metric for quantifying the change in accuracy between two predictive systems in the case of a binary outcome, the InterModel Vigorish (IMV). The IMV is based on an analogy to well-characterized physical systems with tractable probabilities: weighted coins. The IMV is always a statement about the change in fit relative to some baseline---which can be as simple as the prevalence---whereas other metrics are stand-alone measures that need to be further manipulated to yield indices related to differences in fit across models. Moreover, the IMV is consistently interpretable independent of baseline prevalence. We illustrate the flexible properties of this metric in numerous simulations and showcase its flexibility across examples spanning the social, biomedical, and physical sciences. The IMV allows for precise answers to questions about changes in model fit in a variety of settings in a manner that we think will be useful for furthering research with binary outcomes.

Key Idea II: Examine all publicly available CDM data

- Use the ***Item Response Warehouse*** (IRW, Domingue et al., 2025)



Item Response Warehouse

The Item Response Warehouse (IRW) is a collection of open, harmonized item response datasets.

- Open: The data are licensed for reformatting and public reuse. Their provenance is documented and code used to transform original data to meet the IRW data standard is available.
- Harmonized: The data are formatted to a common standard. The goal of the IRW is to make large quantities of data available for standardized analysis given the common formatting.
- Item Response: The IRW is comprised of cross-classified item responses. The key element required of data for inclusion in the IRW is that the data need to report individual responses that are cross-classified by the focus of measurement (typically a person) and the measurement tool (typically an item).

These datasets in IRW span a wide range of measures; examples include measures of ability in education settings, measures of partisanship based on voting records, and personality measures based on survey questions. Responses can be scored in a variety of ways (dichotomous, polytomous, continuous) and may contain additional data (e.g., response time).

- Examine the degree to which IRT vs. CDM —which utilize quite distinctive notions regarding the nature of ability—produce ***different predictions of response behavior in the real-world.***

Agenda

- Motivation: rethinking about comparing IRT vs. CDM
- Key Ideas: focus on out-of-sample predictive accuracy with empirical data
- **CDM Basics**
- Simulations: various conditions when data is generated by CDMs
- Empirical Datasets (N = 9)
- Discussion and future directions
- Q&A

CDM: Q matrix

- Each item requires a specific subset of attributes.

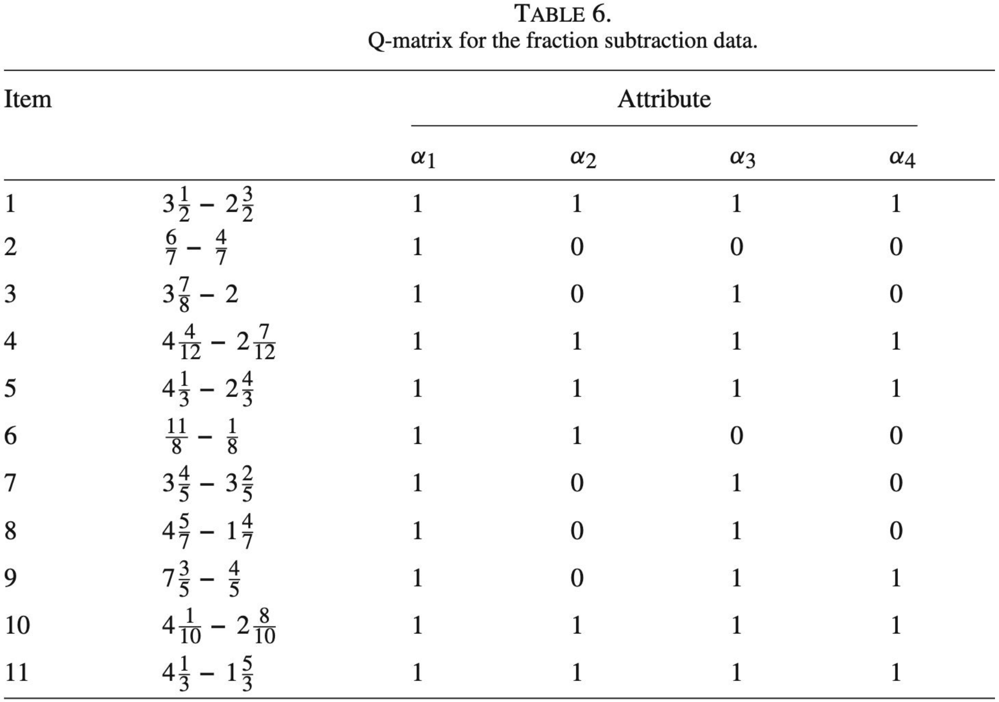


TABLE 6.
Q-matrix for the fraction subtraction data.

Item		Attribute			
		α_1	α_2	α_3	α_4
1	$3\frac{1}{2} - 2\frac{3}{2}$	1	1	1	1
2	$\frac{6}{7} - \frac{4}{7}$	1	0	0	0
3	$3\frac{7}{8} - 2$	1	0	1	0
4	$4\frac{4}{12} - 2\frac{7}{12}$	1	1	1	1
5	$4\frac{1}{3} - 2\frac{4}{3}$	1	1	1	1
6	$\frac{11}{8} - \frac{1}{8}$	1	1	0	0
7	$3\frac{4}{5} - 3\frac{2}{5}$	1	0	1	0
8	$4\frac{5}{7} - 1\frac{4}{7}$	1	0	1	0
9	$7\frac{3}{5} - \frac{4}{5}$	1	0	1	1
10	$4\frac{1}{10} - 2\frac{8}{10}$	1	1	1	1
11	$4\frac{1}{3} - 1\frac{5}{3}$	1	1	1	1

α_1: performing basic fraction subtraction operation

$\alpha_{2:}$ simplifying/reducing

$\alpha_{3:}$ separating whole number from fraction

$\alpha_{4:}$ borrowing one from whole number to fraction

Table adapted from De La Torre & Chiu, 2016, p. 267

CDM: non-compensatory vs. compensatory models

DINA: mastery of all required attributes is necessary for a correct response.

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}. \tag{2}$$

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{3}$$

DINO: mastery of at least one required attribute is sufficient for a correct response.

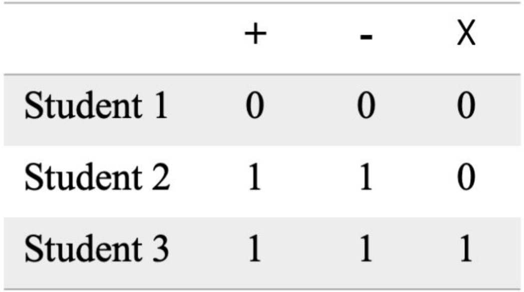
$$\eta_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}. \tag{4}$$

GDINA: offers a flexible generalization of the DINA and DINO models by accommodating both main effects and higher-order interactions among attributes.

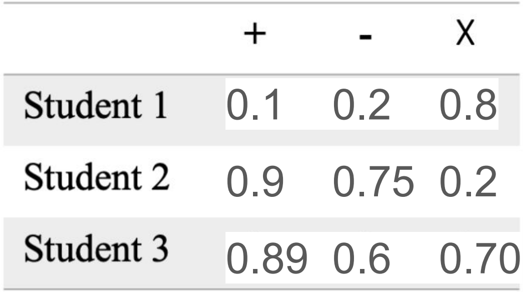
$$P(Y_{ij} = 1 | \boldsymbol{\alpha}_{\ell j}^{*}) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk} \alpha_{\ell jk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'} \alpha_{\ell jk} \alpha_{\ell jk'} + \cdots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{\ell jk} \tag{5}$$

CDM: attribute estimation

- Package ‘GDINA’ (Ma et al., 2025)
- (1) Maximum A Posteriori (**MAP**): a binary vector representing the most likely full mastery pattern across all attributes simultaneously.
- (2) marginal mastery probabilities (**mp**): a vector of probabilities for each attribute, reflecting the confidence of mastery for each skill individually.



	+	-	X
Student 1	0	0	0
Student 2	1	1	0
Student 3	1	1	1



	+	-	X
Student 1	0.1	0.2	0.8
Student 2	0.9	0.75	0.2
Student 3	0.89	0.6	0.70

Agenda

- Motivation: rethinking about comparing IRT vs. CDM
- Key Ideas: focus on out-of-sample predictive accuracy with empirical data
- CDM Basics
- **Simulations: various conditions when data is generated by CDMs**
- Empirical Datasets (N = 9)
- Discussion and future directions
- Q&A

Simulation 1: Data generated via the hierarchical CDM with varying attribute correlations

DGM: A hierarchical CDM where the attributes are generated from the multivariate normal threshold model (Chiu et al., 2009):

- We vary the within-person **attribute correlation parameter ρ** from Unif(0, 0.8).
- Q-matrix (30 items, 5 attributes) from the GDINA package (Ma et al., 2025).
- **Sample size**: 200, 500, 1000
- **CDM Models**: DINA vs. GDINA
- Use the ground-true probabilities to generate new responses for out-of-sample prediction.

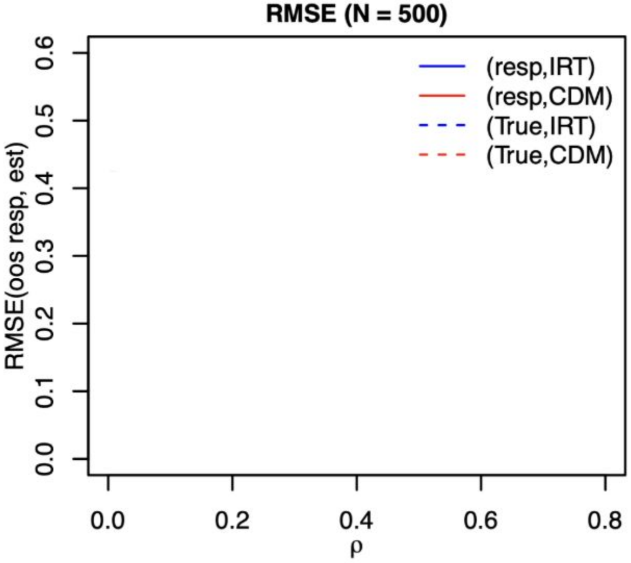
DAM: CDM vs. 2PL

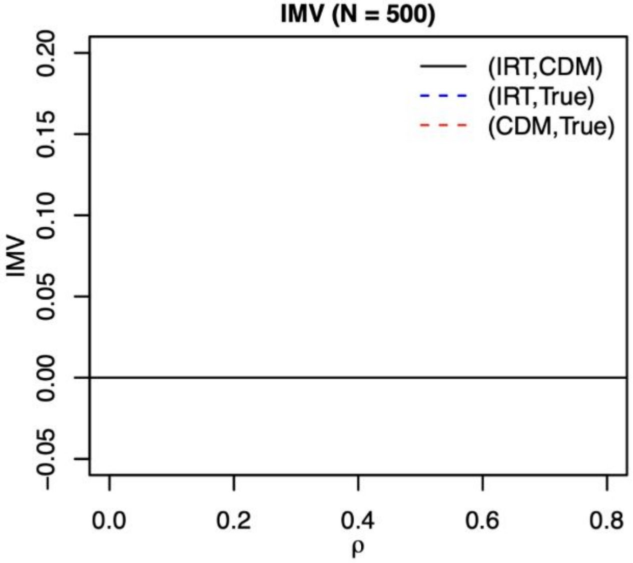
- **person attribute estimator**:

  maximum a posteriori (MAP) vs. marginal mastery probabilities (mp)

Simulation 1: Data generated via the hierarchical CDM with varying attribute correlations

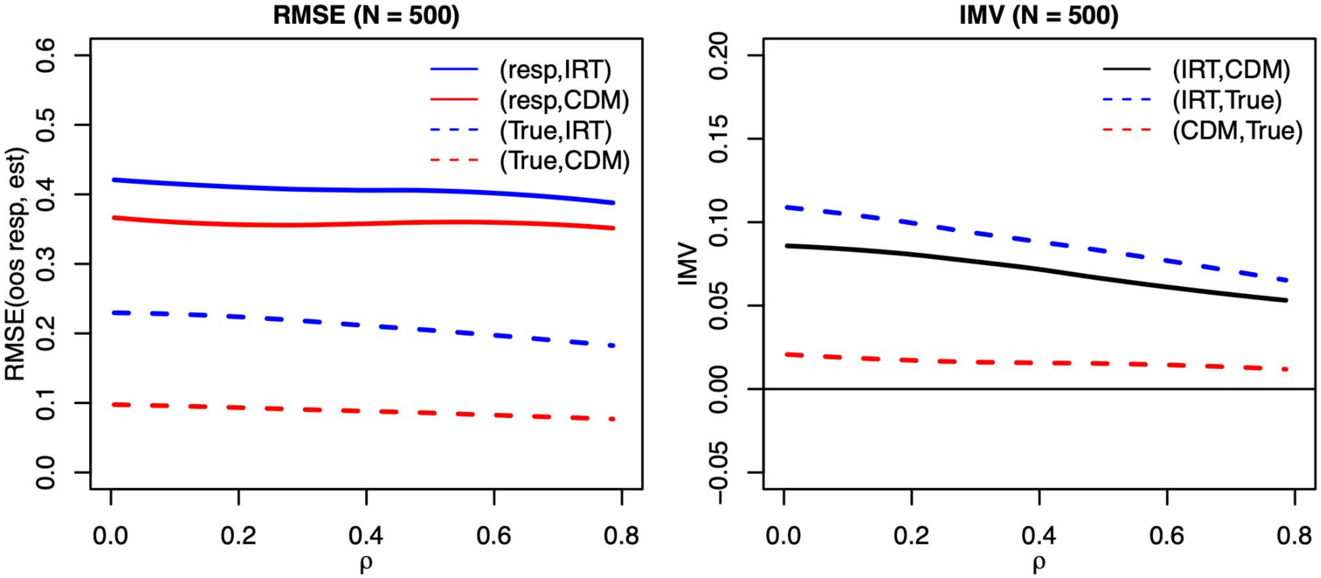
- Data generated from DINA





Simulation 1: Data generated via the hierarchical CDM with varying attribute correlations

- Data generated from DINA

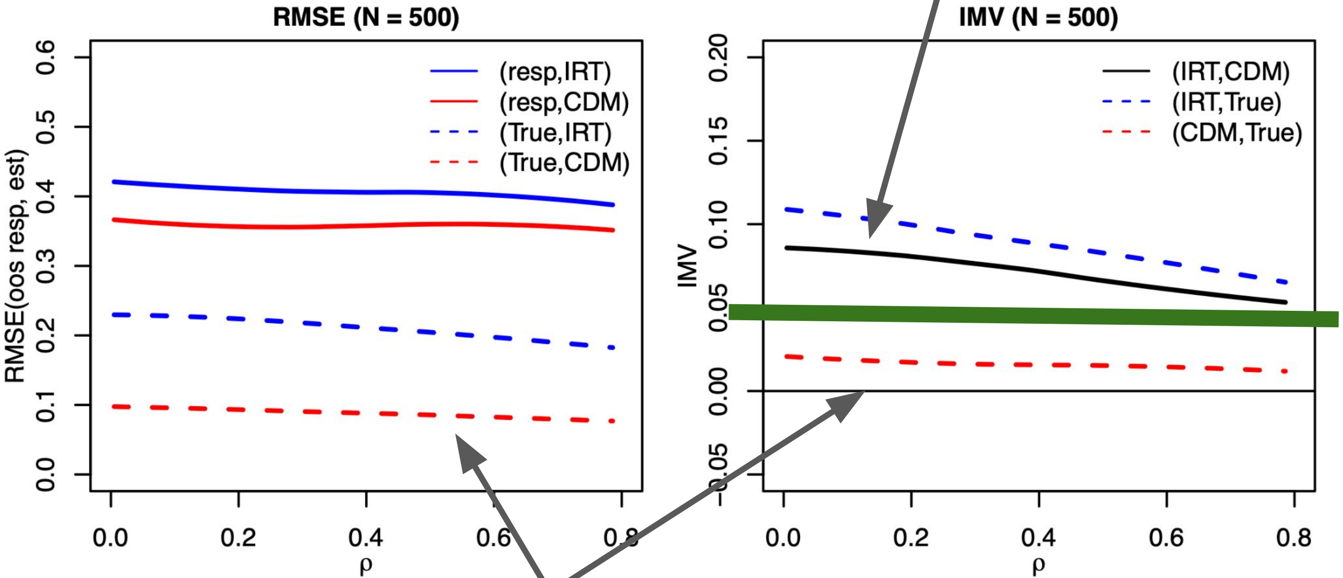


Simulation 1: Data generated via the
with varying attribute correlations

- Data generated from DINA

- CDM estimates are consistently high-quality

- IRT estimates improve as attributes are more correlated.



The solid black line **IMV(IRT, CDM)** compares observable quantities → key quantities in work with empirical data

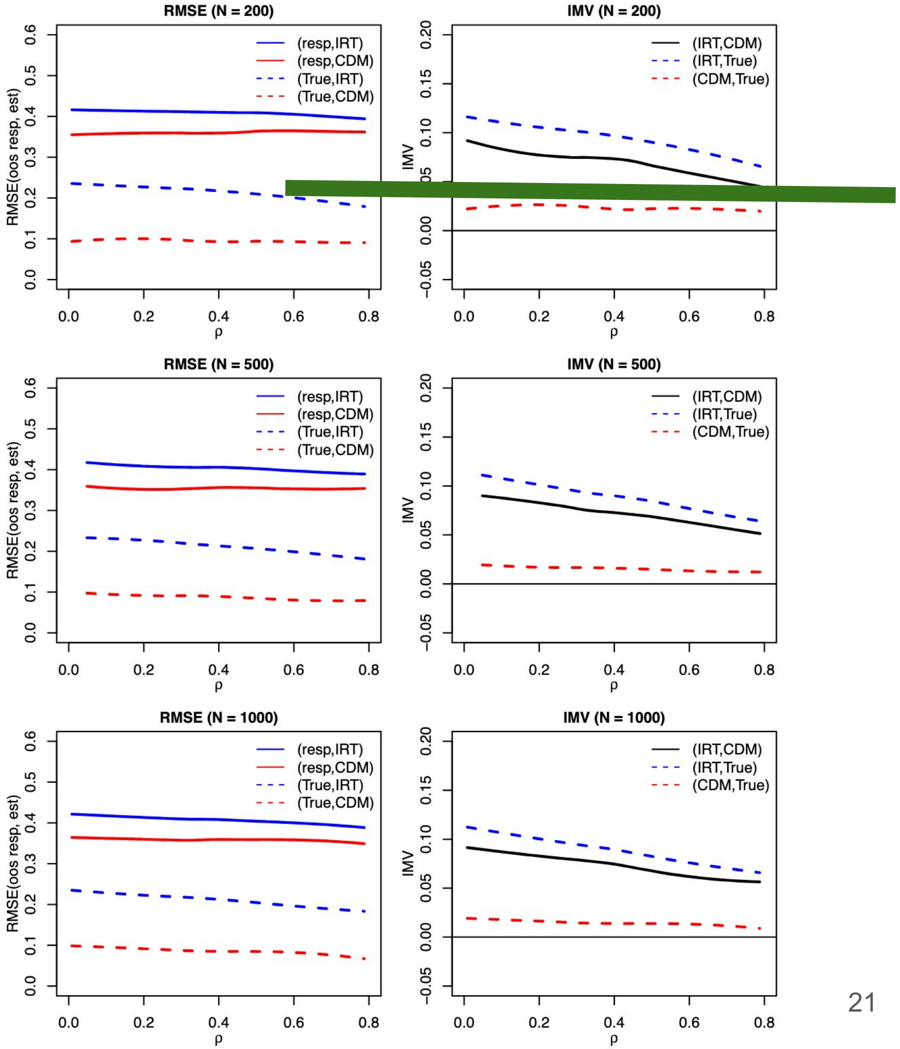
The dashed lines involve the true response probabilities.

Benchmark: IMV(IRT, CDM) > 0.05

More from Simulation 1:

- Increases in sample size lead to smaller gains in IMV(2PL, CDM).

- **But the benchmark value of 0.05 still holds.**

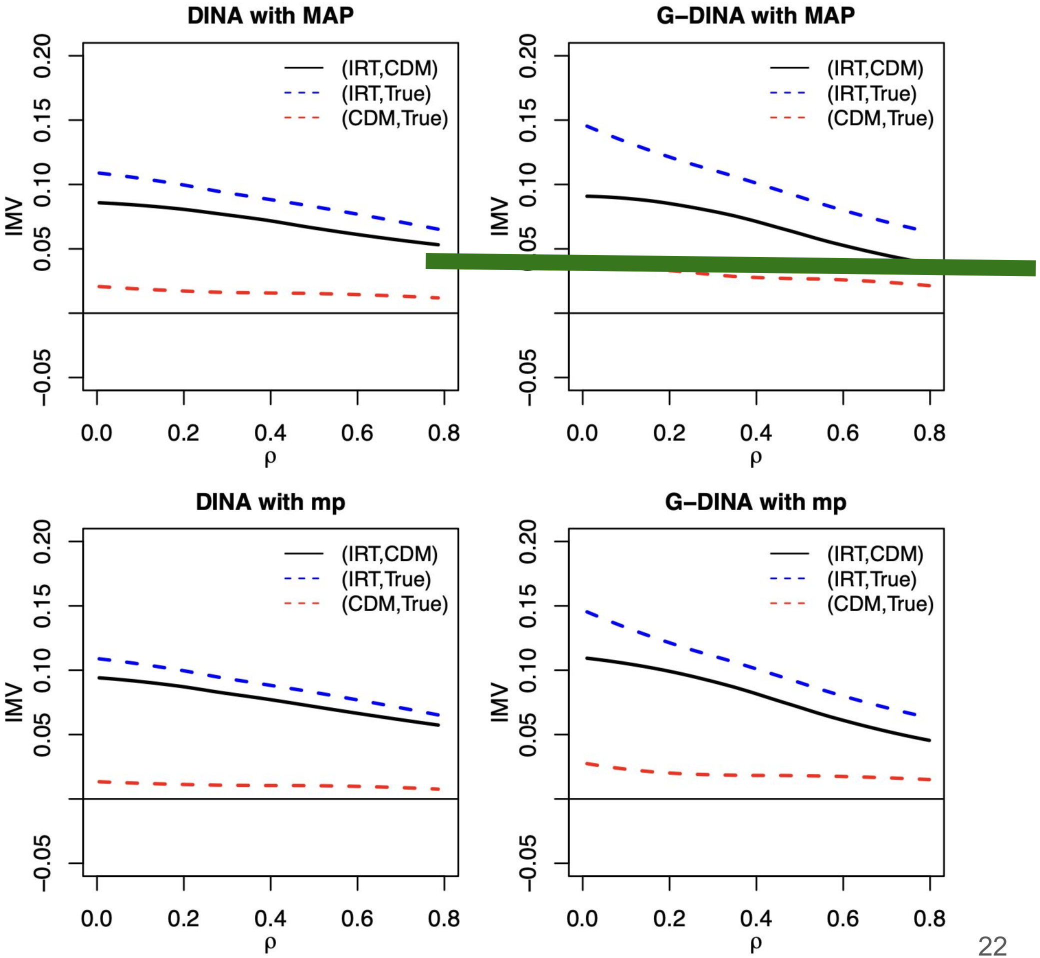


More from Simulation 1:

- The choice of attribute estimator yields minimal differences in IMV.

- Both DINA and GDINA share the same story about the results above.

- **The benchmark value of 0.05 still holds.**

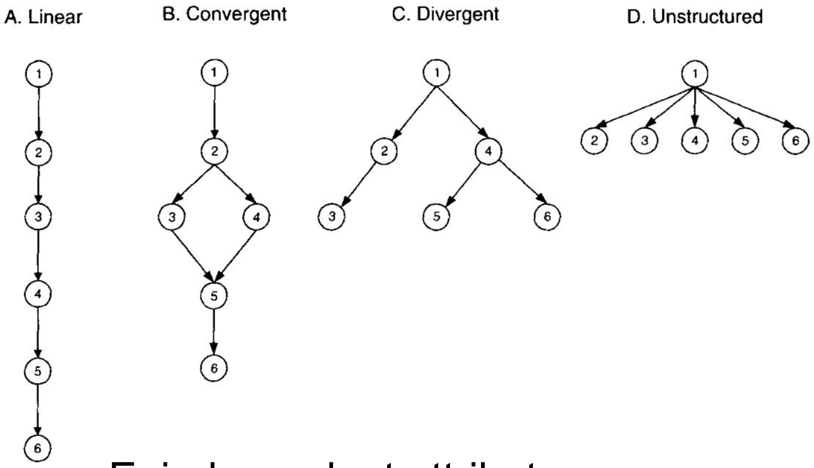


Simulation 2: Data generated via a CDM with different hierarchical attribute structures

- Different **attribute hierarchies** may lead to different model fit (Leighton et al., 2004; J. Templin & Bradshaw, 2014; Wang & Lu, 2021).

- Simulation 1: re-generate responses from the true model probabilities.

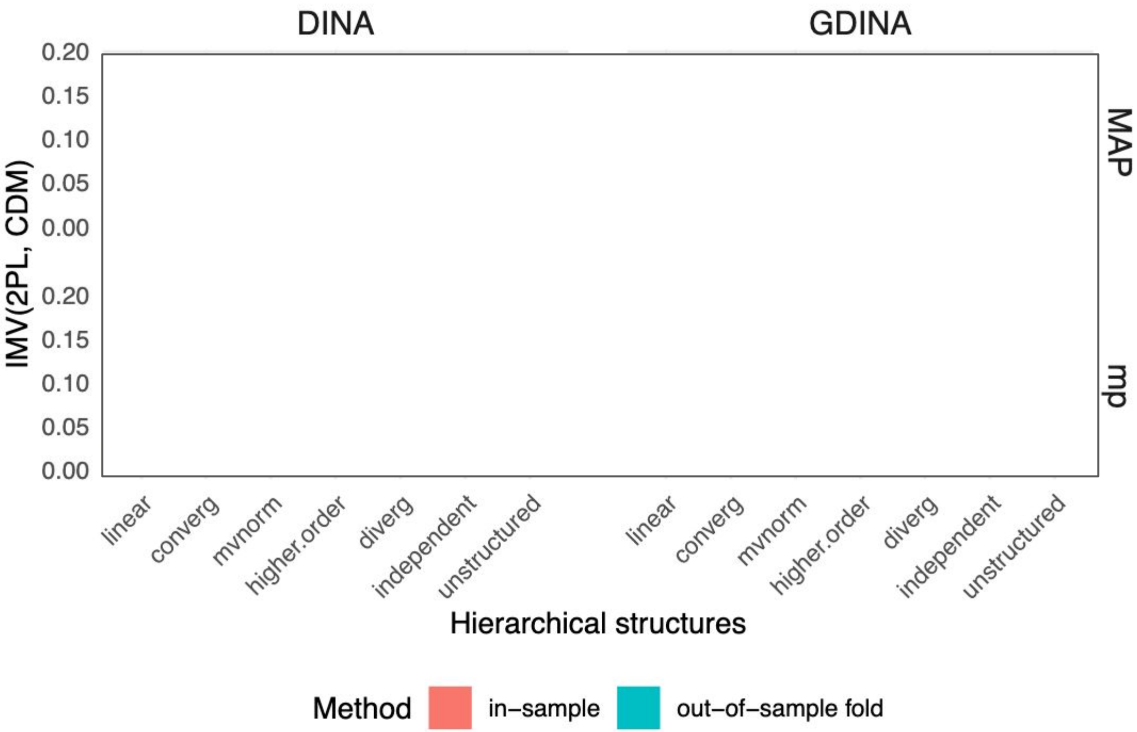
- Simulation 2: apply **5-fold** out-of-sample prediction → more precise benchmark in-sample and out-of-sample IMV values.



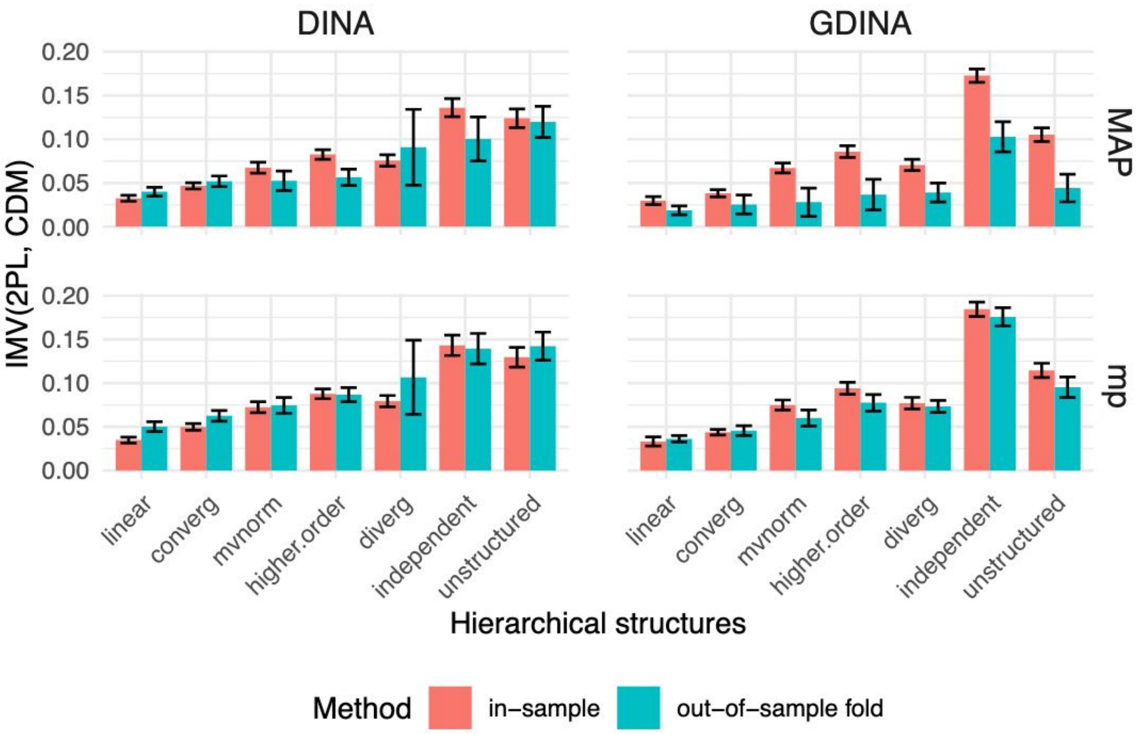
E. independent attributes
F. a higher-order structure
(a multivariate normal distribution with attribute correlation ρ = 0.5)

Figure adapted from Leighton et al., 2014.

Simulation 2: Data generated via a CDM with different hierarchical attribute structures

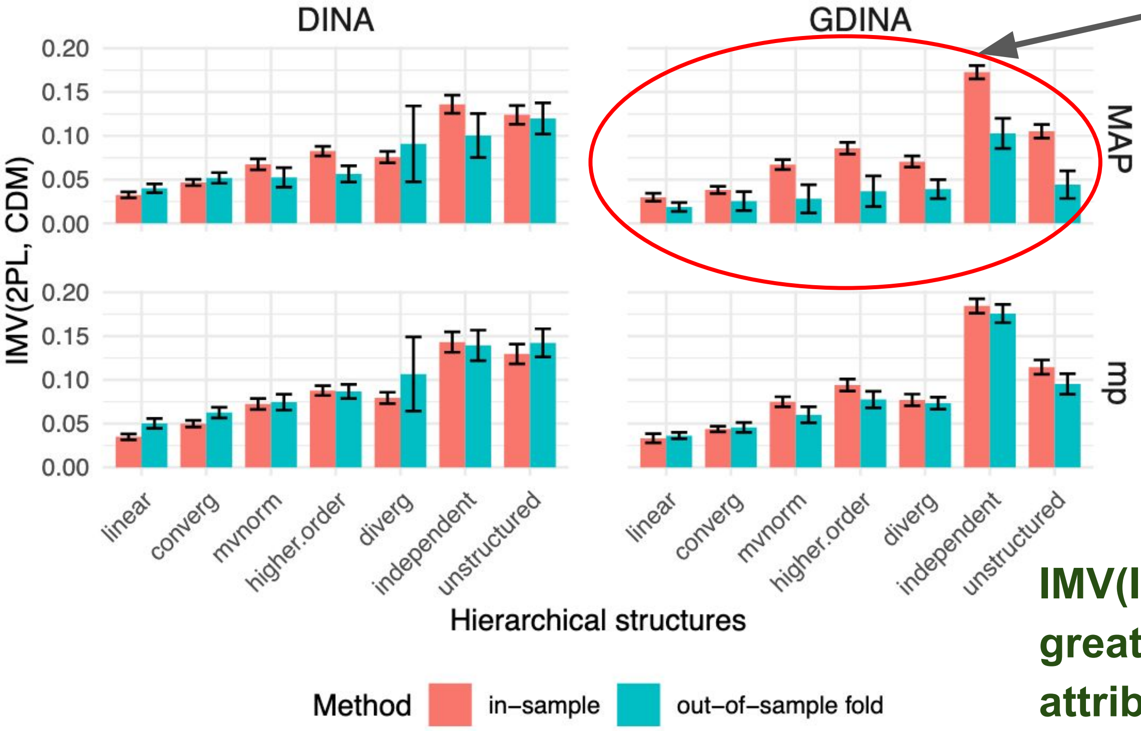


Simulation 2: Data generated via a CDM with different hierarchical attribute structures



Simulation 2: Data generated via a CDM with different hierarchical attribute structures

Evidence of overfitting.



IMV(IRT,CDM) values are greater than 0.02 across all attribute structures.

Take-away from the simulations:

- Model recovery is consistently superior under a CDM when the data are generated from class-based latent traits, as compared to a conventional IRT model predicated on a continuously varying latent trait.

- Benchmark: IMV(IRT, CDM) is between [0.02, 0.05]

- The potential risk of overfitting when using MAP estimates for the G-DINA model.

Agenda

- Motivation: rethinking about comparing IRT vs. CDM
- Key Ideas: focus on out-of-sample predictive accuracy with empirical data
- CDM Basics
- Simulations: various conditions when data is generated by CDMs
- **Empirical Datasets (N = 9)**
- Discussion and future directions
- Q&A

Real World Data!

Empirical Study

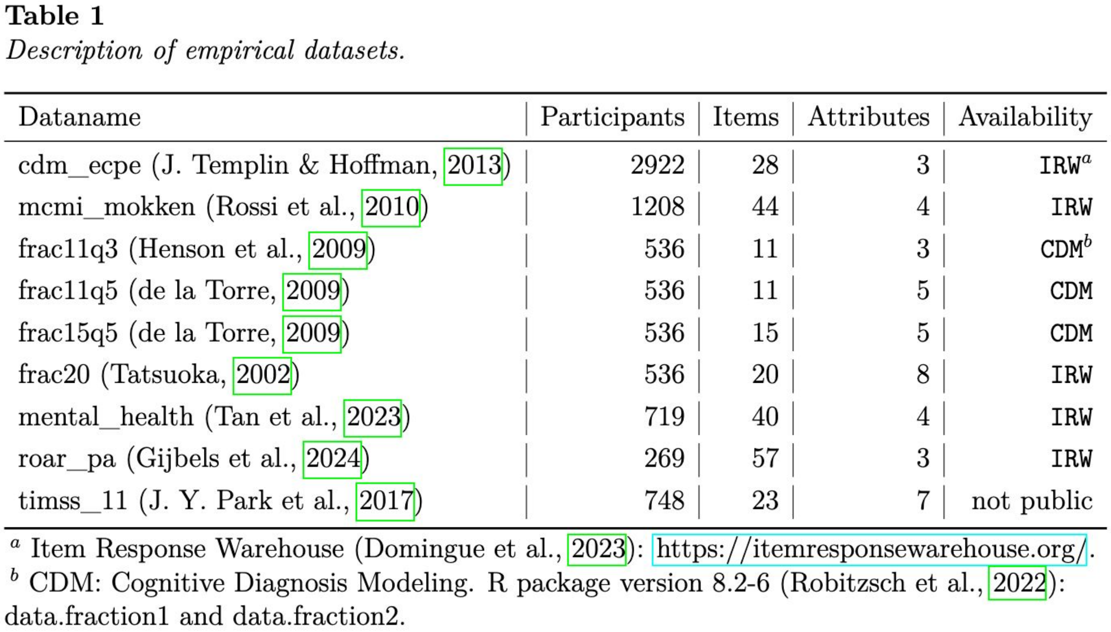


Table 1

Description of empirical datasets.

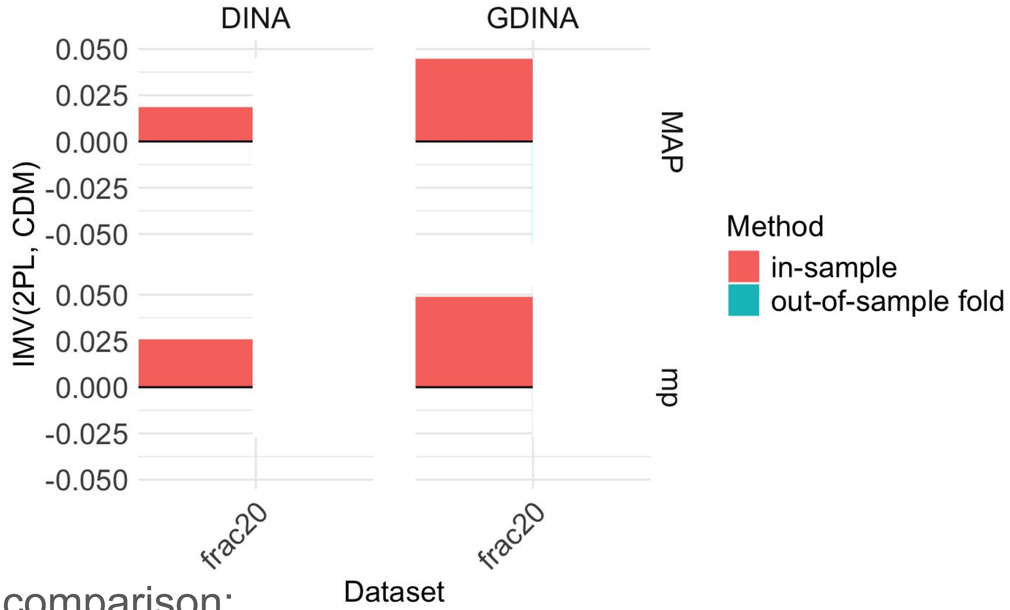
Dataname	Participants	Items	Attributes	Availability
cdm_ecpe (J. Templin & Hoffman, 2013)	2922	28	3	IRW[a]
mcmi_mokken (Rossi et al., 2010)	1208	44	4	IRW
frac11q3 (Henson et al., 2009)	536	11	3	CDM[b]
frac11q5 (de la Torre, 2009)	536	11	5	CDM
frac15q5 (de la Torre, 2009)	536	15	5	CDM
frac20 (Tatsuoka, 2002)	536	20	8	IRW
mental_health (Tan et al., 2023)	719	40	4	IRW
roar_pa (Gijbels et al., 2024)	269	57	3	IRW
timss_11 (J. Y. Park et al., 2017)	748	23	7	not public

[a] Item Response Warehouse (Domingue et al., 2023): https://itemresponsewarehouse.org/.

[b] CDM: Cognitive Diagnosis Modeling. R package version 8.2-6 (Robitzsch et al., 2022): data.fraction1 and data.fraction2.

- Data: N = 9 datasets has an existing skill attribution classification.
- Analysis Plan: **IMV(2PL, CDM)**, vary by models (DINA vs. G-DINA) and by estimation methods (MAP vs. mp)
- **We hypothesize that the CDM will provide qualitatively superior predictive accuracy.**

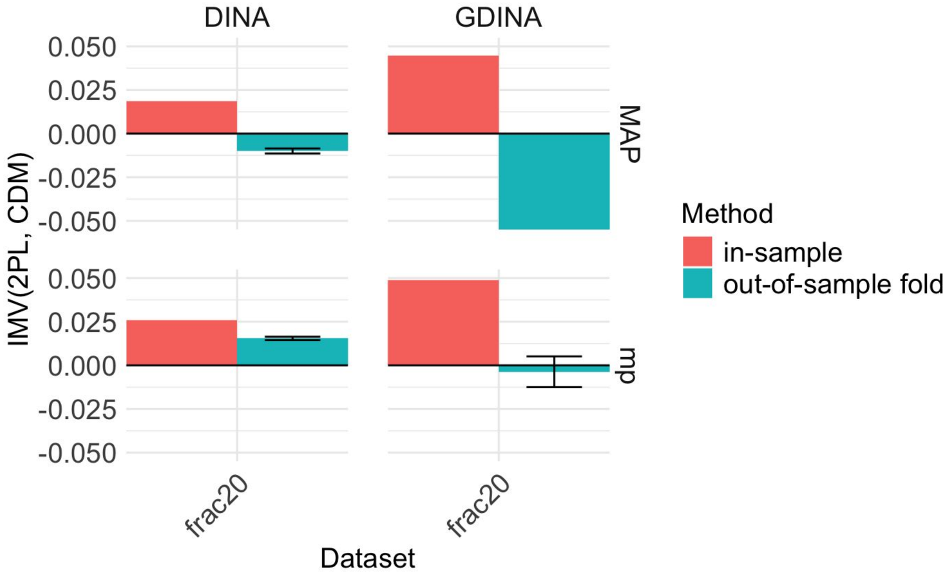
Empirical Results: Let’s take a look at Fraction20 first!



Story from the **in-sample** comparison:

- CDMs fit better than the 2PL model.
- The G-DINA model fits better than the DINA model
- MAP and mp produce comparable results.

Empirical Results: Let’s take a look at Fraction20 first!



In fact:

- There is an overfitting issue, especially for the G-DINA model.
- MAP estimator underperforms relative to the mp estimator.

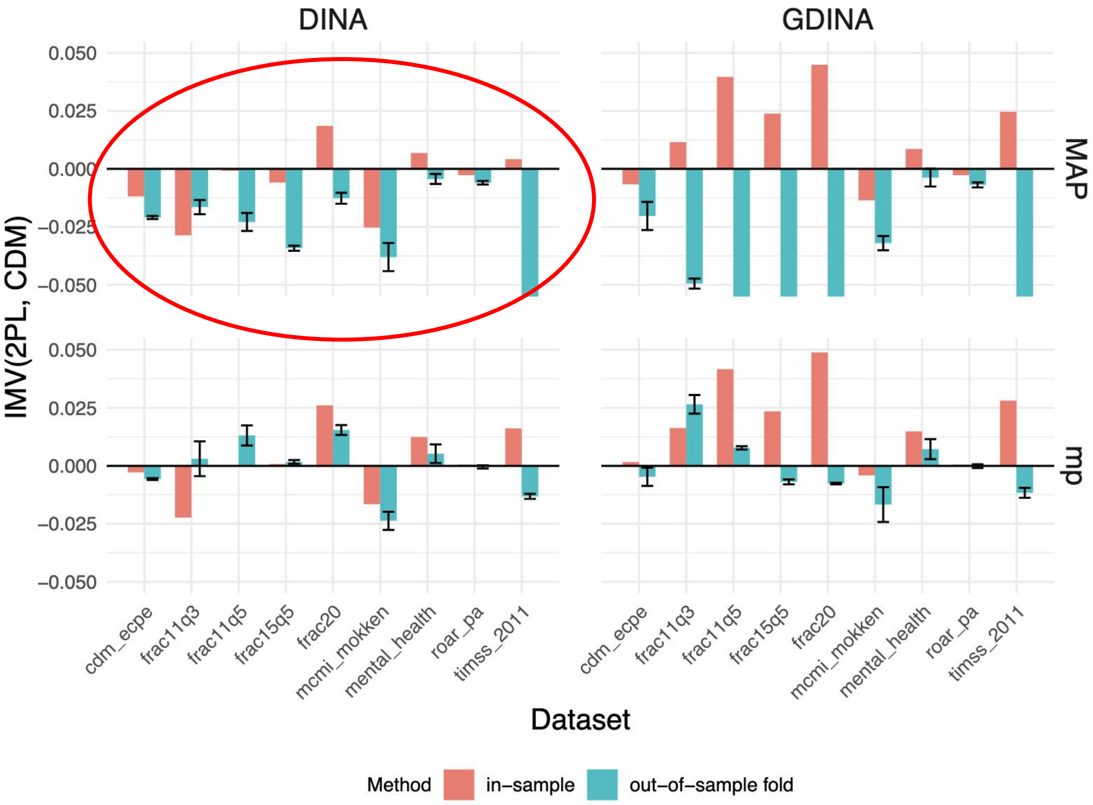
Benchmark from simulation:
IMV(IRT, CDM) > 0.02

Empirically, IRT models generally outperform CDM models.

- Overfitting of the CDM model is a pervasive concern.

- Moving away from a probabilistic representation of attribute mastery negatively impacts predictive accuracy.

- Concern of “retrofitting” is real, but maybe it is more than that …



Agenda

- Motivation: rethinking about comparing IRT vs. CDM
- Key Ideas: focus on out-of-sample predictive accuracy with empirical data
- CDM Basics
- Simulations: various conditions when data is generated by CDMs
- Empirical Datasets (N = 9)
- **Discussion and future directions**
- Q&A

Conclusion & Implication

- **Predictive accuracy** as a central model evaluation criterion, extending beyond traditional in-sample fit indices.
- The out-of-sample approach reveals the true story: the risk of outfitting.
- With the IRW (a wider range of empirical data) → comprehensive view

Is this the end of story for CDM?

- Researchers and practitioners may need to balance the diagnostic appeal of CDMs with the fact that their complexity can come at the cost of predictive accuracy.

“Brainstorming” future directions

- Bring more “original CDM data” into this game?

- Q-matrix validation (Ma et al., 2020, Nájera et al., 2021)

- Find the alternative? multidimensional-IRT?

Broader Discussion:

1. **How do we think about balancing the inferior performance of the model fitting vs. the potential appeal?**

   - “Question is weather the misfitting model still provides a useful scale, e.g. having predictive validity (school, job, therapy).” — Klaas Sijtsma, BEAR Seminar, Fall 2021

2. **How can we develop assessments that provide actionable, diagnostic information without relying on CDMs?**

   - ordered multiple-choice items (Briggs et al., 2006)

   - developmental assessment (Wilson, 2008)

Thank you!

PREPRINT

Wanjing Anya Ma

wanjingm@stanford.edu

https://anyawma.github.io/

Stanford | GRADUATE SCHOOL OF EDUCATION

Selected References

Ma, W., Minchen, N., & de la Torre, J. (2020). Choosing between CDM and unidimensional IRT: The proportional reasoning test case. *Measurement: Interdisciplinary Research and Perspectives*, *18*(2), 87-96.

Yamaguchi, K., & Okada, K. (2020). Variational Bayes inference for the DINA model. *Journal of Educational and Behavioral Statistics*, *45*(5), 569-597.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317-339.

von Davier, M. (2014). The log‐linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series*, *2014*(2), 1-13.

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional ‘diagnostic’classification models—a commentary. *Psychometrika*, *79*(2), 340-346.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12 (6), 1100–1122.

Domingue, B. W., Rahal, C., Faul, J., Freese, J., Kanopka, K., Rigos, A., ... & Tripathi, A. S. (2025). The InterModel Vigorish (IMV) as a flexible and portable approach for quantifying predictive accuracy with binary outcomes. *PloS one*, *20*(3), e0316491.

Domingue, B. W., Kanopka, K., Kapoor, R., Pohl, S., Chalmers, R. P., Rahal, C., & Rhemtulla, M. (2024). The intermodel vigorish as a lens for understanding (and quantifying) the value of item response models for dichotomously coded items. *psychometrika*, *89*(3), 1034-1054.

Domingue, B. W., Braginsky, M., Caffrey-Maffei, L., Gilbert, J. B., Kanopka, K., Kapoor, R., Lee, H., Liu, Y., Nadela, S., Pan, G., Zhang, L., Zhang, S., & Frank, M. C. (2025). An introduction to the Item Response Warehouse (IRW): A resource for enhancing data usage in psychometrics. *Behavior research methods*, *57*(10), 276. https://doi.org/10.3758/s13428-025-02796-y.