AI and Cognitive Diagnostic Models for Educational Assessment:

Connections, Integrations, and New Directions

Richard M. Golden
School of Behavioral and Brain Sciences
University of Texas at Dallas
golden@utdallas.edu

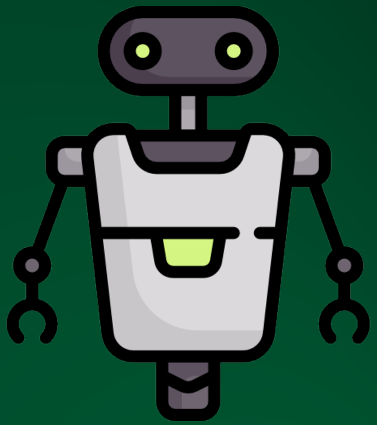
COINS
Cognitive Informatics & Statistics
LAB

Golden, R. M. (2025, April 26). *AI and Cognitive Diagnostic Models for Educational Assessment: Connections, Integrations, and New Directions.* [NCME organized session]. Bridging theory and Practice: Fostering innovation in diagnostic measurement To enhance education. National Council on Measurement in Education, Denver Colorado.

DENVER
2025 ANNUAL MEETING
ncme
APRIL 23-26, 2025

Some Benefits of Artificial Intelligence for Education Assessment

- Automated Essay and Constructed Response Data Scoring
  (Flor and Hao, 2021)
- Adaptive Personalized Learning/Intelligent Tutoring Systems
  (Almond, 2007; Green et al., 2011; Reichenberg, 2018; Madison and Bradshaw, 2018; Levy, 2019)
- Automated Item Generation
  (including programming problems and items with images)
  (Lauc et al., 2016; Song and Du, 2025; Popescu et al., 2023)



Challenges in using AI for Assessment

- Assessment using Testable, Transparent Explanatory Models
- Assessment results/explanations should (ideally) be unique
- Assessment Task Invariance
- Assessment Diverse Group Invariance
- Diverse Subpopulation Predictive Model Fit
- Causal Model Fit

Assessment as Evidential Reasoning

(Messick, 1994; Mislevy et al., 2006)

STEP 1: CONSTRUCTS:

What skills or other attributes should be assessed and WHY?

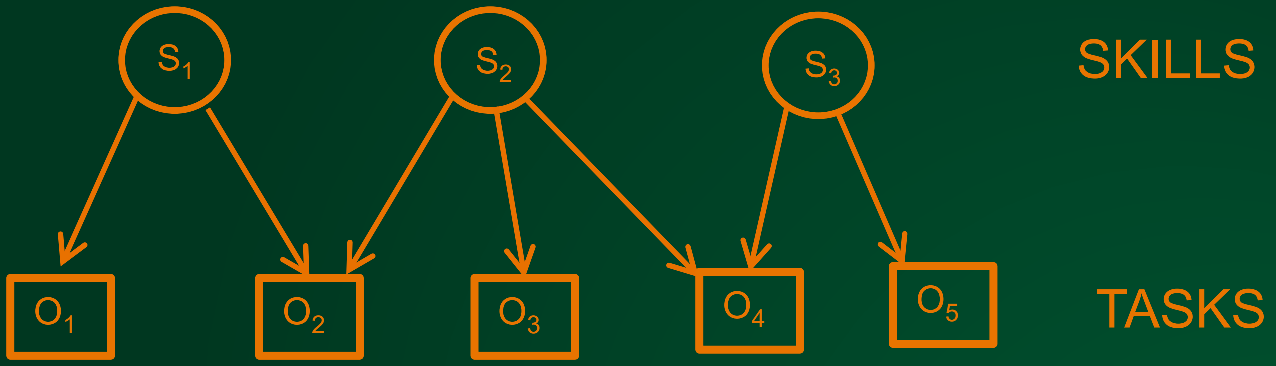
STEP 2: TASKS:

What behaviors or performances reveal those constructs?

STEP 3: CONNECTIONS: How are Tasks and Constructs related?

This methodology implies the existence of a causal chain of evidence which can be inspected, tested, and documented.

Cognitive Diagnostic Models



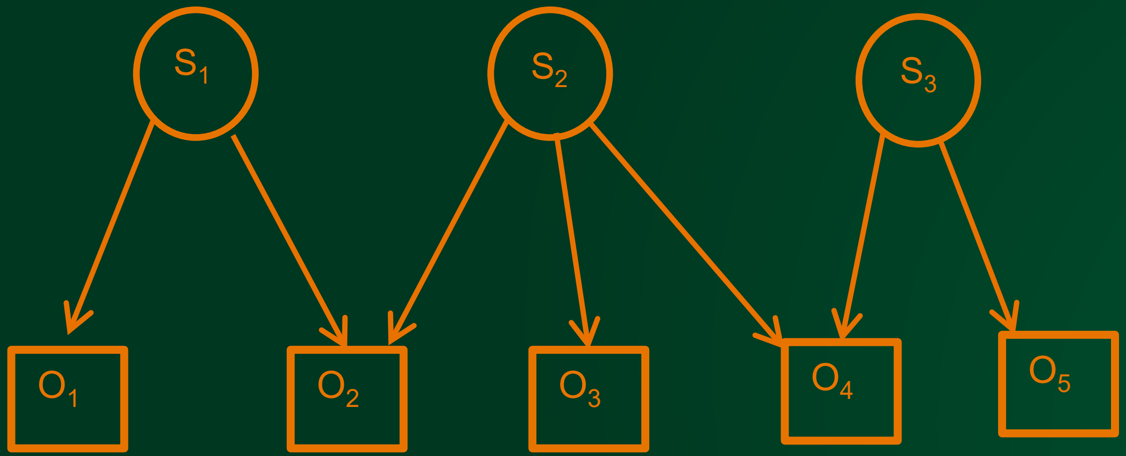
Proficiency Model: joint distribution of skills (model of student population)

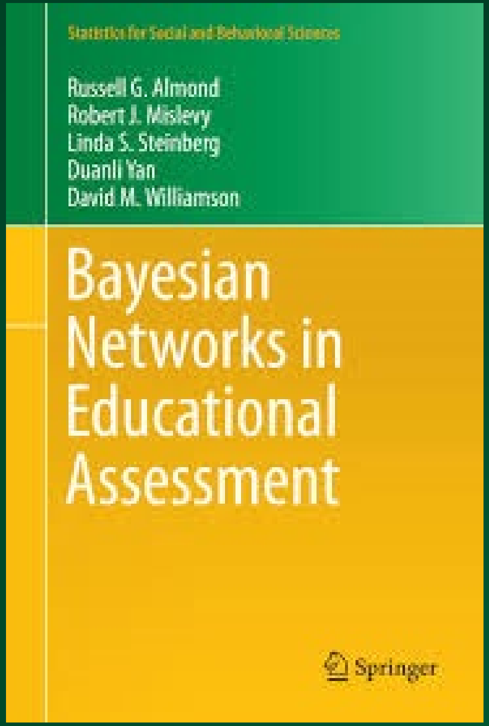
Evidence Model: conditional distribution of responses given skills (model of items)

Inference: Given behavioral observables O_1 O_2 O_3 O_4 O_5 compute probable values of S_1 S_2 S_3

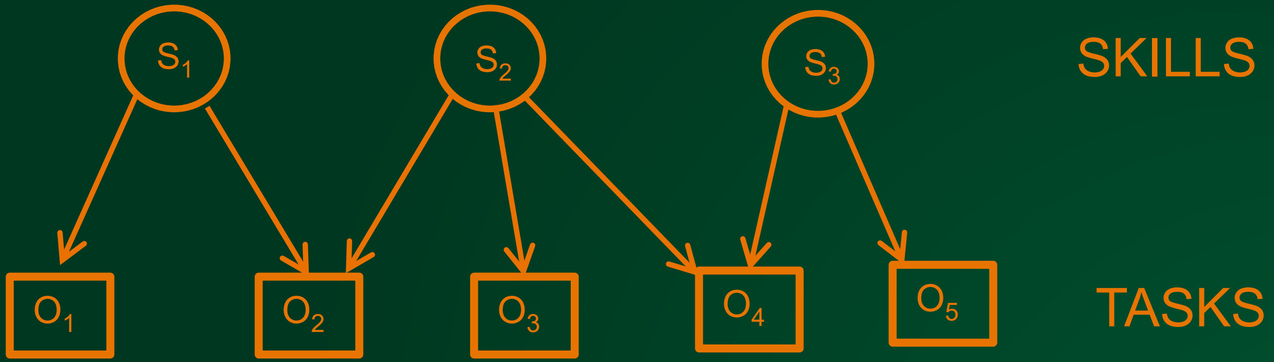
Supports: Form Invariance, Detection of DIF, Assessment through Evidence based reasoning

A Cognitive Diagnostic Model is an example of the well-studied Bayesian Network in AI



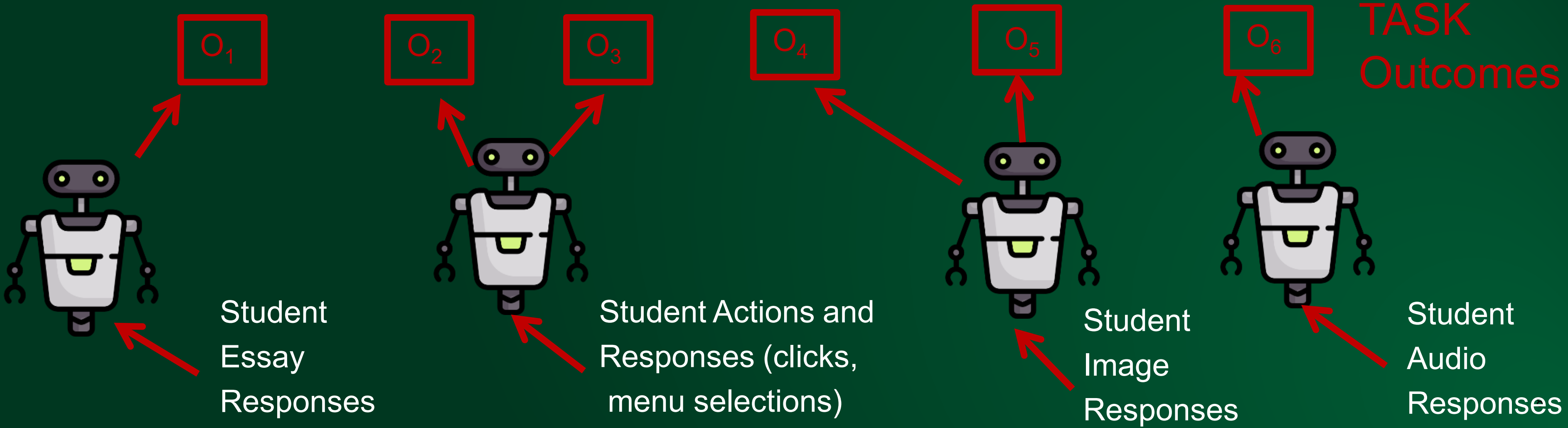


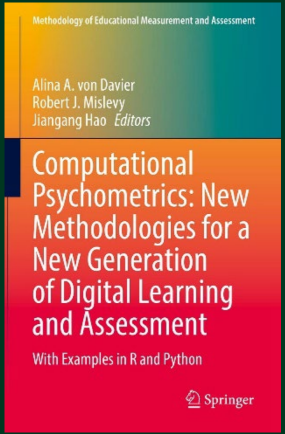
Cognitive Diagnostic Models: Applications



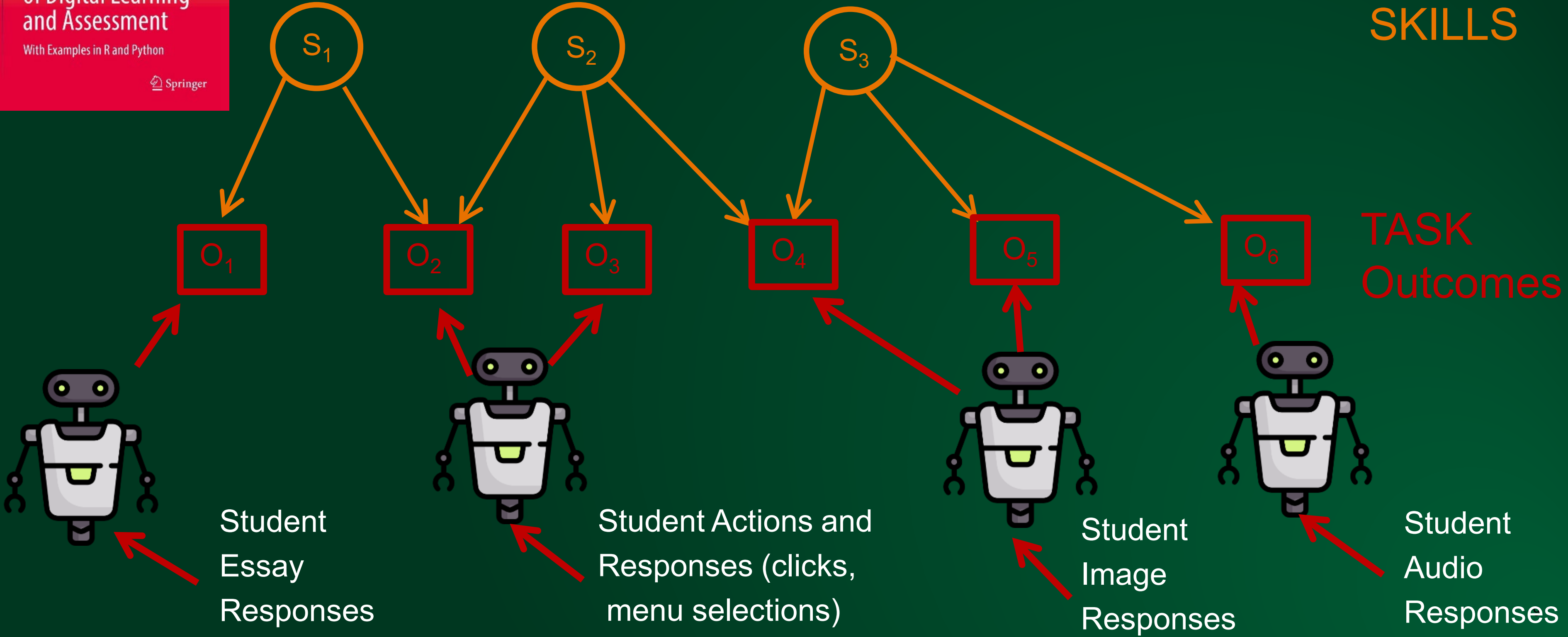
- Navvy Assessments (Laine Bradshaw)
- Dynamic Learning Maps (Jake Thompson)
- Education Companion App by ACTNext (Deonovic et al., 2019)
- Recent Reviews (Deonovic et al., 2019; Williamson, 2023; Wang et al., 2024)
- Evaluating LLMs and other AI systems using CDM methods

Typical Deep Learning Network in AI is a Predictive (not Explanative) Learning Machine



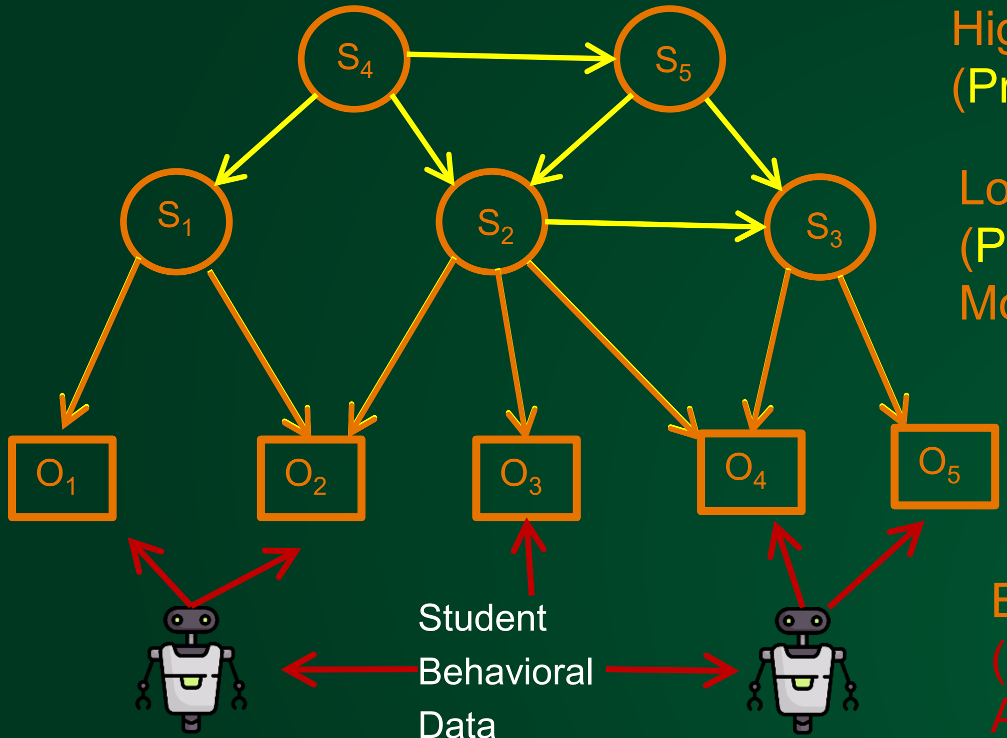


Cognitive Diagnostic Models: Evidence Identification



Cognitive Diagnostic Models: Deep Proficiency Models

(Almond et al., 2015; Kim et al., 2016;Gu, 2024)



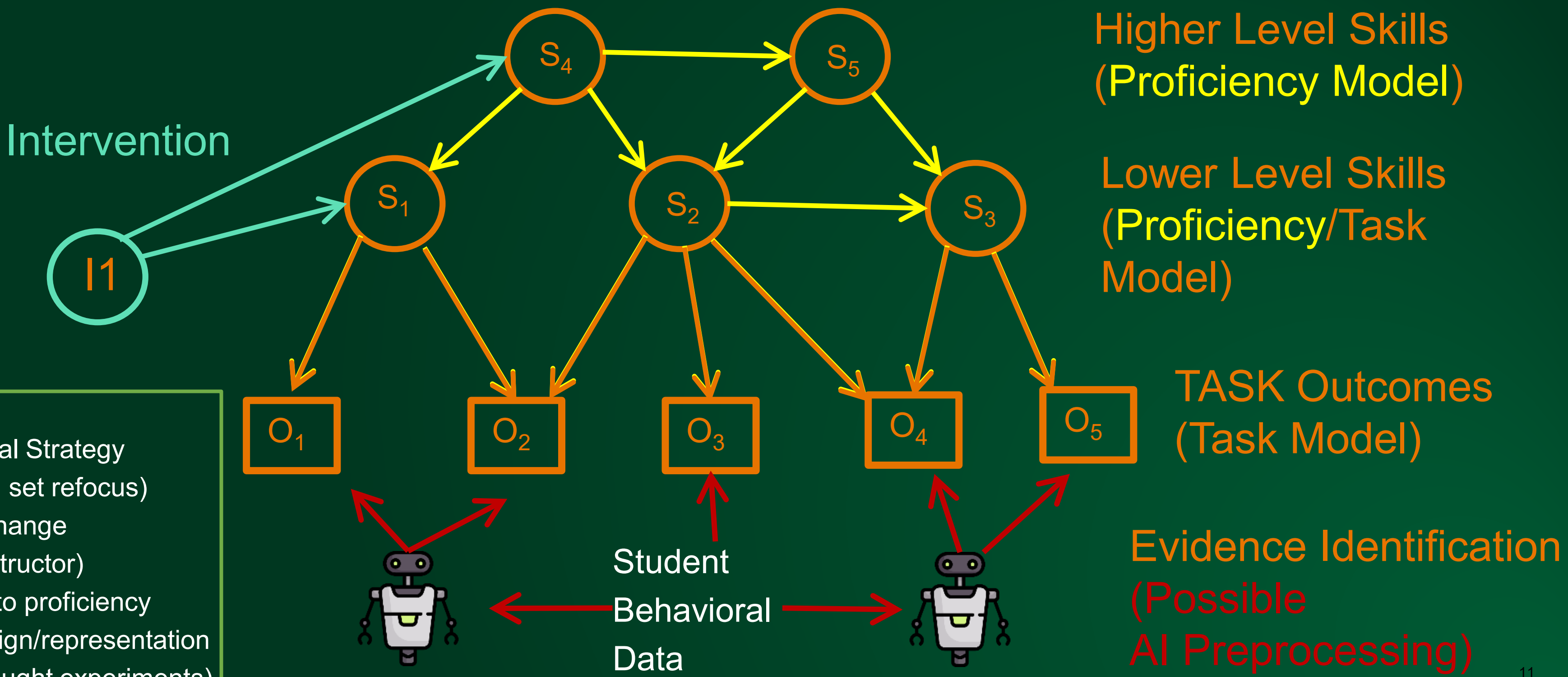
Higher Level Skills (Proficiency Model)

Lower Level Skills (Proficiency/Task Model)

TASK Outcomes (Task Model)

Evidence Identification (Possible AI Preprocessing)

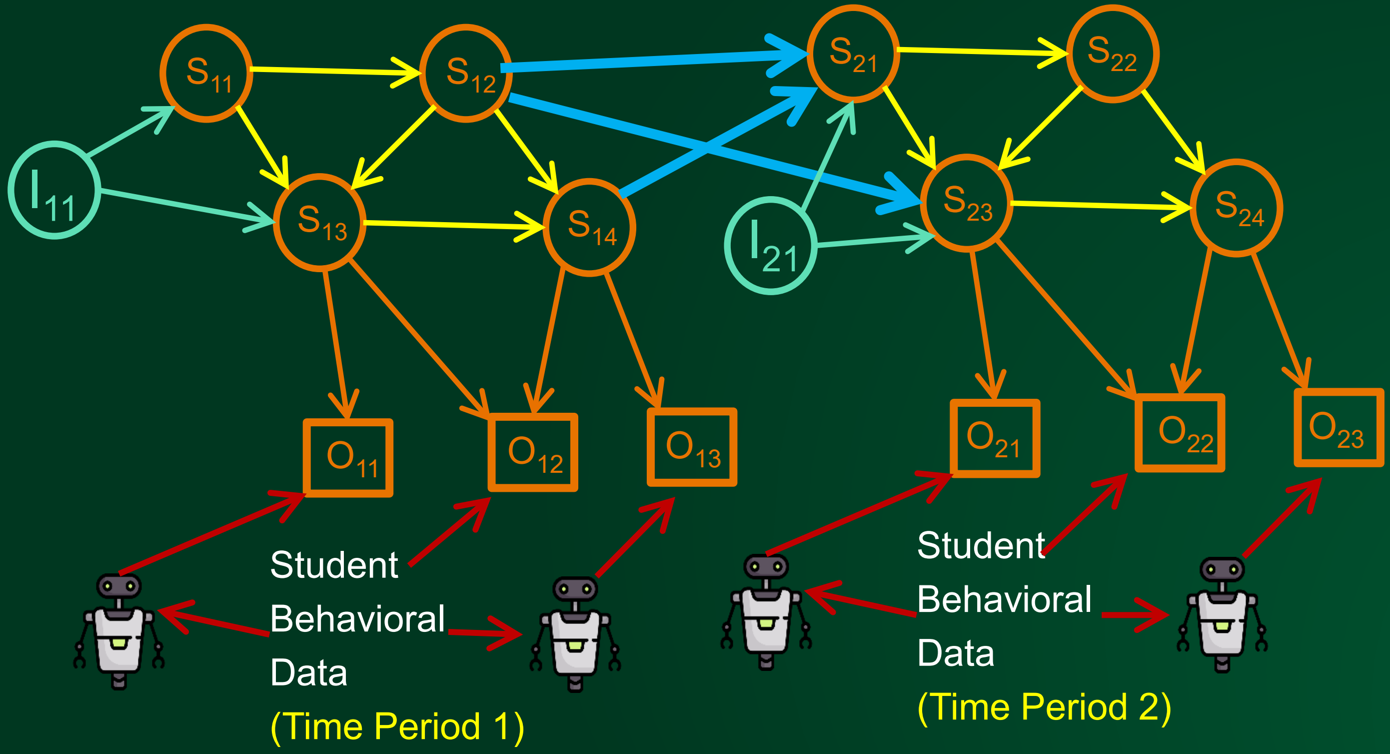
Cognitive Diagnostic Models:
Deep Proficiency Models+Intervention



Interventions
- Instructional Strategy
  (e.g., skill set refocus)
- Context Change
  (e.g., Instructor)
- Insights into proficiency model design/representation
  (e.g., thought experiments)

Cognitive Diagnostic Models: Dynamic Bayes Net

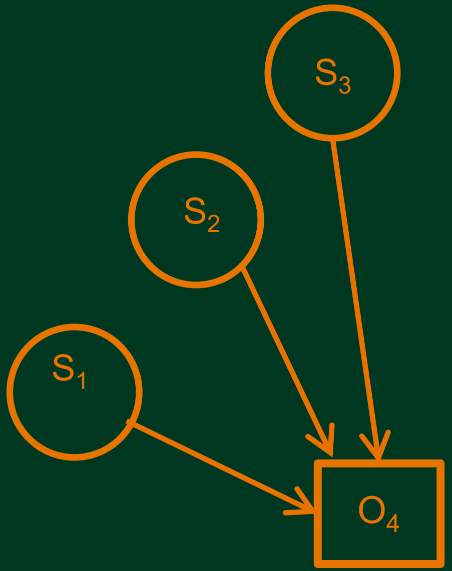
(Almond, 2007; Green et al., 2011; Reichenberg, 2018; Madison and Bradshaw, 2018; Levy, 2019)



This type of Architecture is Common in the AI Reinforcement Learning Literature
It is used for Planning and Control (Golden, 2020, Ch.1 and Ch.12)

This architecture can be used to analyze and design curricula And AI-driven personalized Learning environments (e.g., Intelligent Tutoring systems)

Reducing Parameter Redundancy to Support Assessment Reliability
(Almond et al., 2001; Levy & Mislevy, 2004; Bradshaw & Templin, 2014)



$$p\left(O_4=1|S_1,S_2,S_3\right)=\frac{1}{1+exp\left(-a_4\theta+b_4\right)}$$

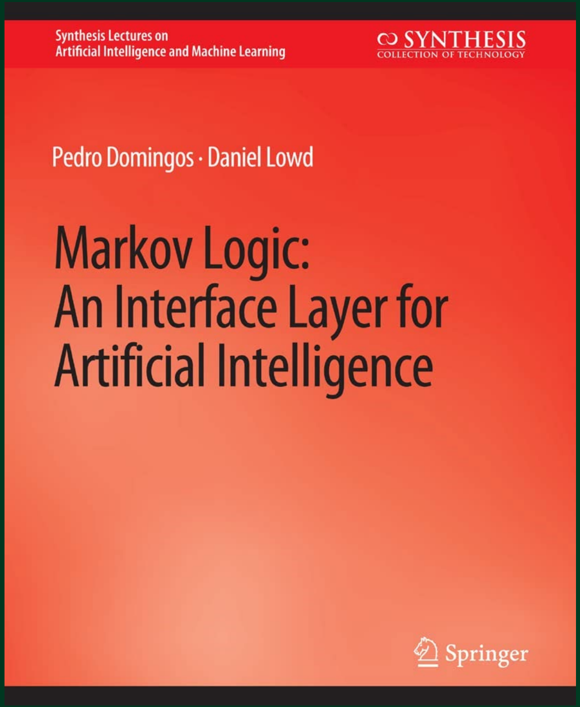
$$\theta=\sum_{j=1}^{3}\beta_j S_j+\beta_{1,2}S_1S_2+\beta_{1,3}S_1S_3+\beta_{2,3}S_2S_3+\beta_{1,2,3}S_1S_2S_3 \quad GDINA$$

$$\theta=S_1\vee S_2\vee S_3 \quad DINO$$

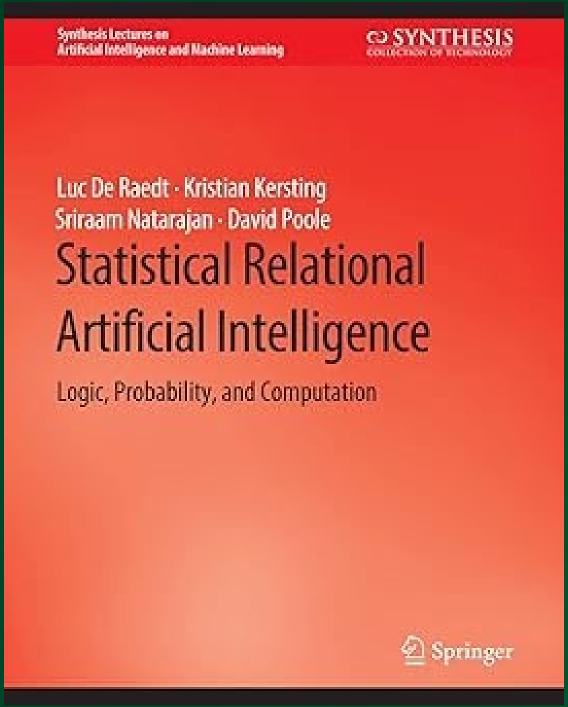
$$\theta=S_1\wedge S_2\wedge S_3 \quad DINA$$

$$\theta=\left(S_1\wedge S_2\right)\vee S_3 \quad MARKOV\ LOGIC\ NET$$

Markov Logic Nets in AI Literature



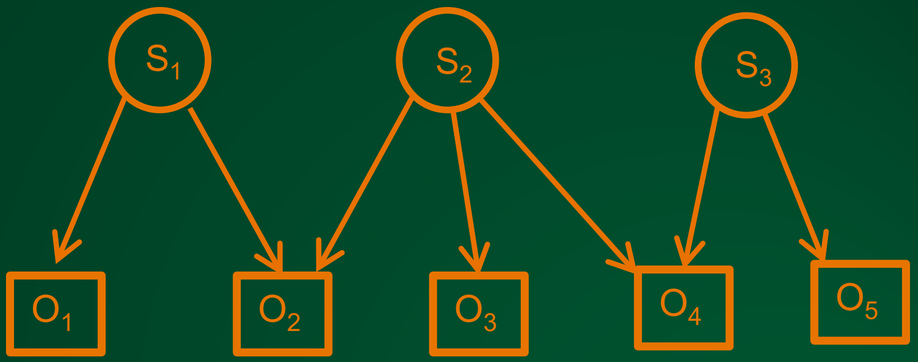
2009



2016

Dealing with Estimation and Inference in Possible Presence of Model Misspecification, Parameter Redundancy, and Empirical Non-Identifiability
(Golden, 2024, under review)

- CDM with many tasks
- Misspecification possible if just 1 task model skill specification is wrong
- Non-identifiability/Parameter Redundancy if just 1 task model too easy or difficult
- Not desirable to reparameterize by deleting task models or linking parameters across task models



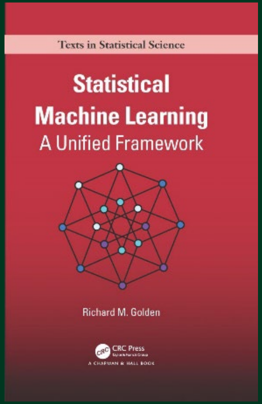
Estimation And Inference In the Presence of Possible Model Misspecification and Parameter Redundancy (Golden, under review)
Technical Report: /https://labs.utdallas.edu/coinslab/files/2024/08/PsychoRedundancyV13_arxivstatTH-1.pdf

Dealing with Inevitable Presence of Redundancy, Non-Identifiability, and Model Misspecification

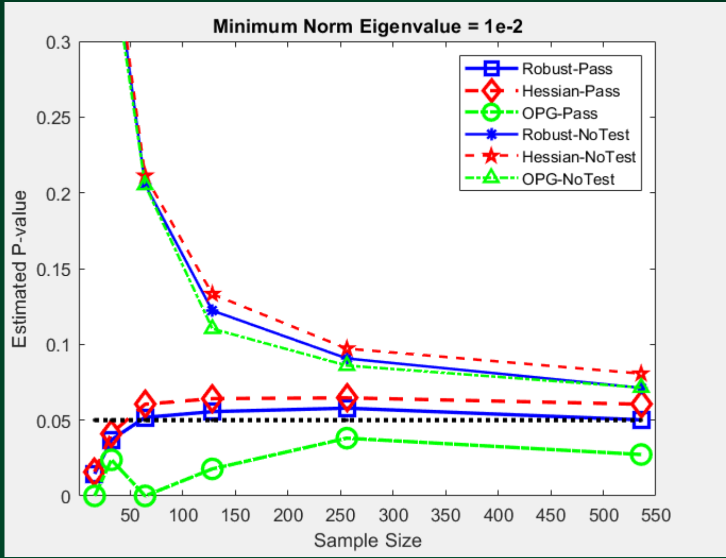
Golden (2024, under review)

- Assume parameters locally identifiable asymptotically
- Non-identifiability for finite samples is ok
- Parameter redundancy for finite samples is ok
- Model may be possibly misspecified
- THEN:

  Theorem shows how to identify permissible inferences (e.g., confidence intervals on individual parameters) and how they can be computed



Proof method involves modifying Huber (1967) analysis of asymptotic distribution of M-estimators (see Golden, 2020, Ch. 15)



Simulation Study showing estimated P-value as a function of sample size for a 30 parameter CDM. New method effective for smaller sample sizes unlike standard inference methods.

Assessment as Evidential Reasoning

(Messick, 1994; Mislevy, 2021)

STEP 1: CONSTRUCTS:

What skills or other attributes should be assessed and WHY?

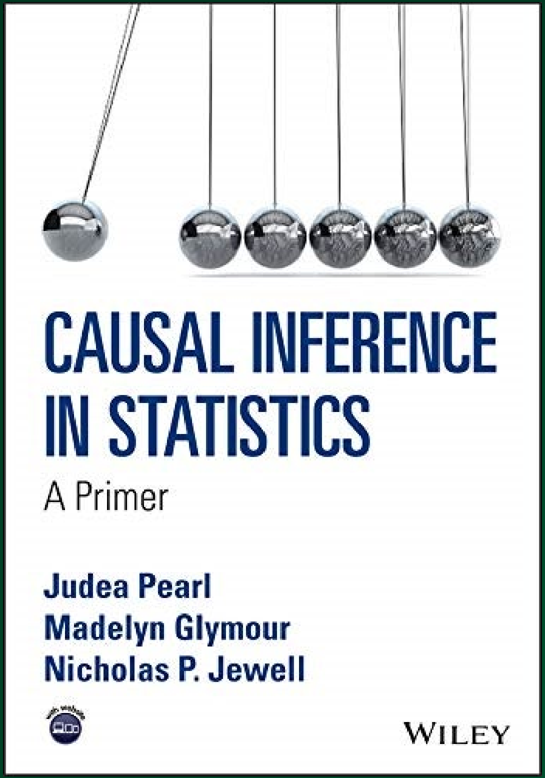
STEP 2: TASKS:

What behaviors or performances reveal those constructs?

STEP 3: CONNECTIONS: How are Tasks and Constructs related?

INTERVENTIONS: (1) Test the theory, and (2) Exploitthe theory

Causal Modeling has been around for decades but only in past five or ten years has it become a hot area in Artificial Intelligence

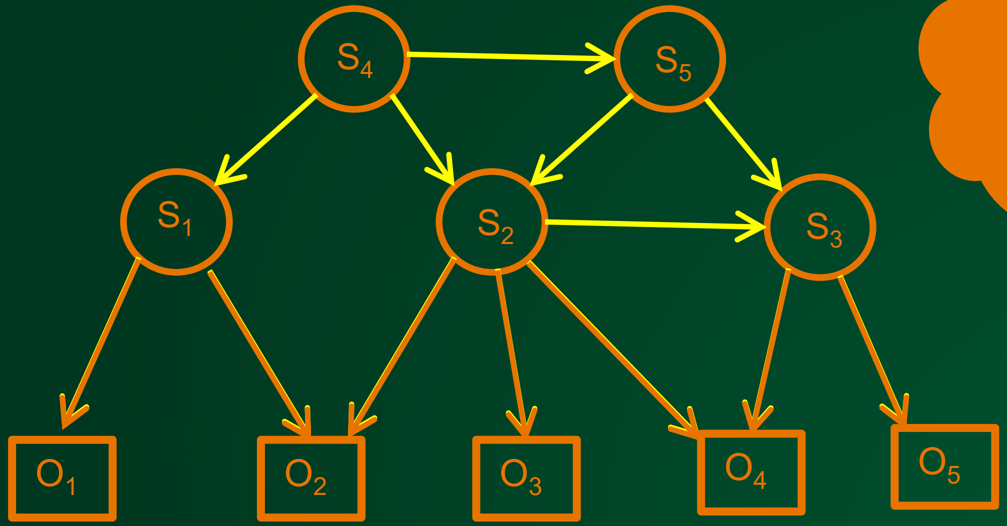


Golden, R. M. *Inferring Causal Relationships from Observational Data Sets*. Talk presented at the Frontiers in Brain Health Lecture Series, Center for Brain Health, University of Texas at Dallas, Dallas, Texas, May 19, 2023. Click here to watch the lecture! (https://www.youtube.com/watch?v=lewmQtYfsSY&list=PLlVrJgppjfmJ0v_-iak0vNFOCLU8A4Nlh&index=2)

It is expected that causal modeling will play an increasingly important role in our cognitive diagnostic models of assessment

Technical (Mathematical) Definition:

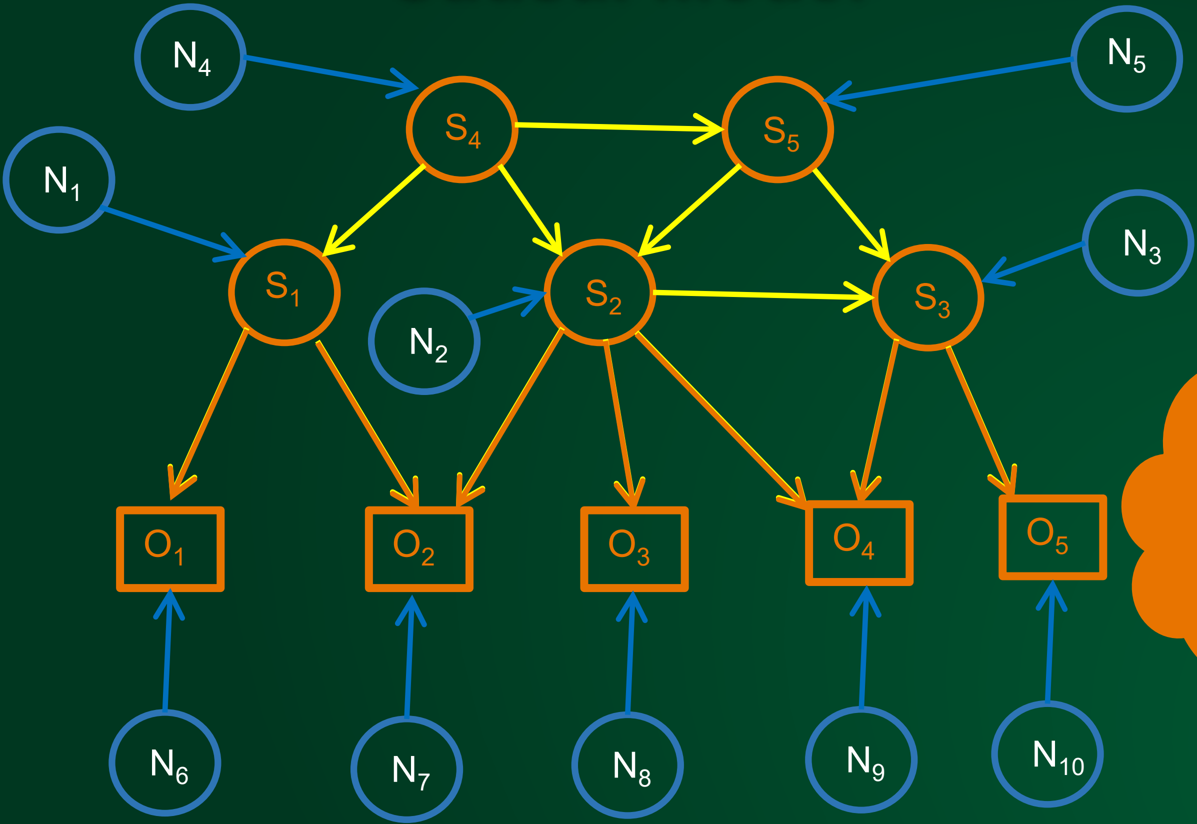
Bayes Net Probability Model



Arrows specify Conditional Independence Assumptions

Technical (Mathematical) Definition:

Causal Model



Arrows specify Causal Modeling Assumptions

Confounder Problem in Causal Inference

Student Motivation	Reading Performance
3	8
0	4
1	2
2	3
1	4

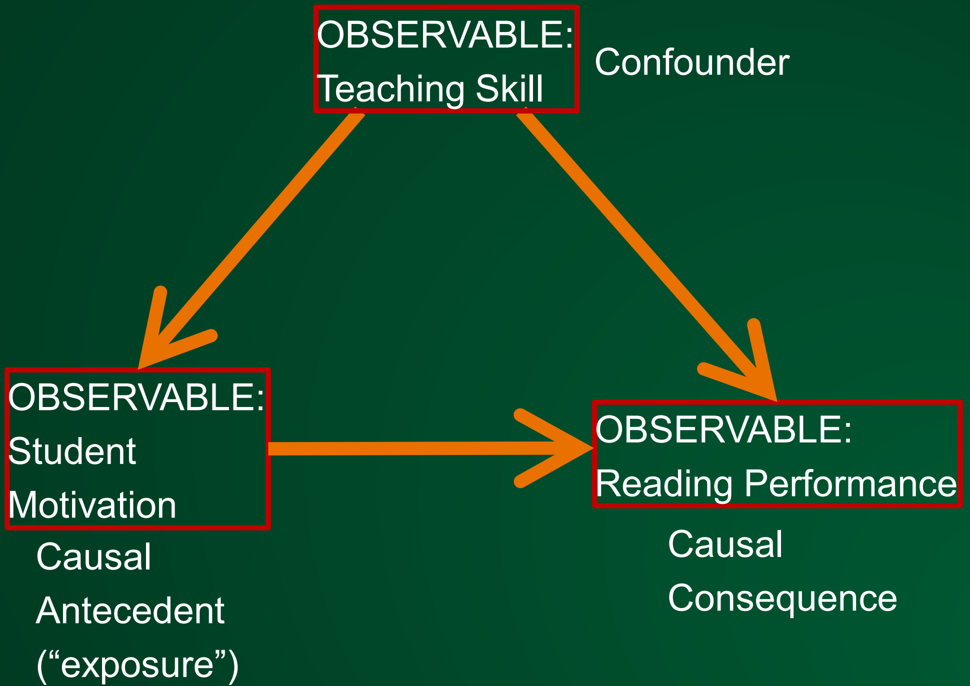
$\boxed{\text{OBSERVABLE: Student Motivation}} \longrightarrow \boxed{\text{OBSERVABLE: Reading Performance}}$

Causal Antecedent ("exposure")

Causal Consequence

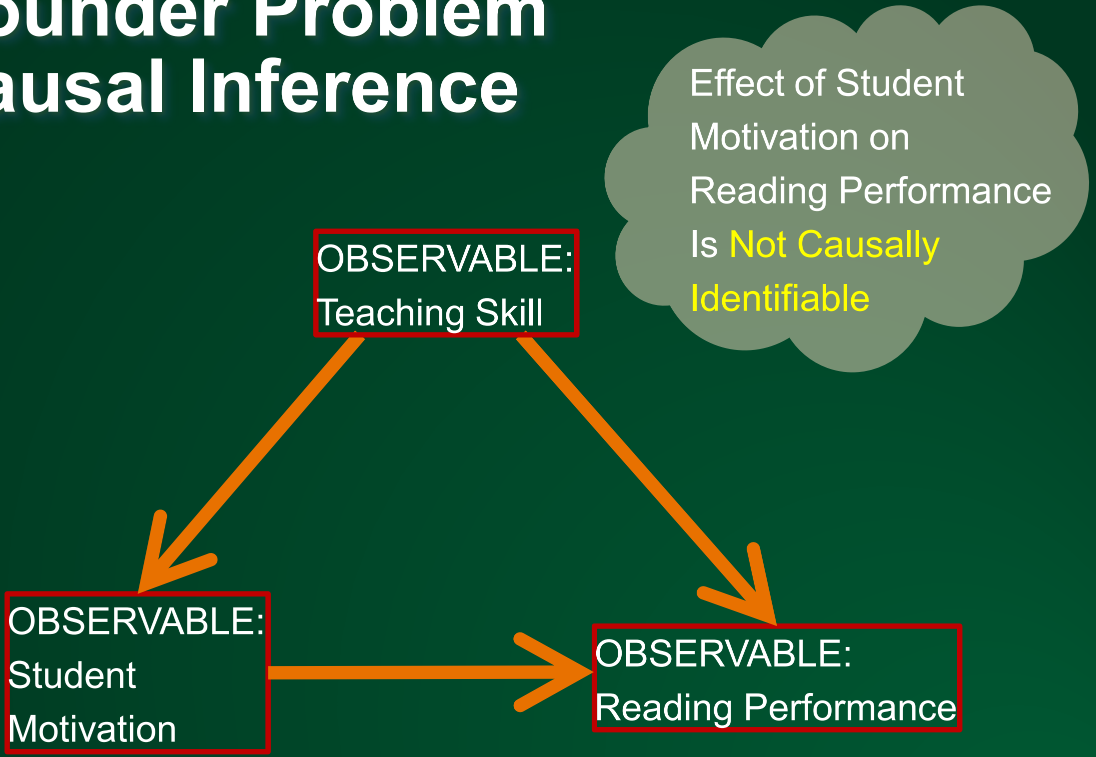
Confounder Problem in Causal Inference

Student Motivation	Reading Performance	Teaching Skill
3	8	3
0	4	0
1	2	0
2	3	5
1	4	1



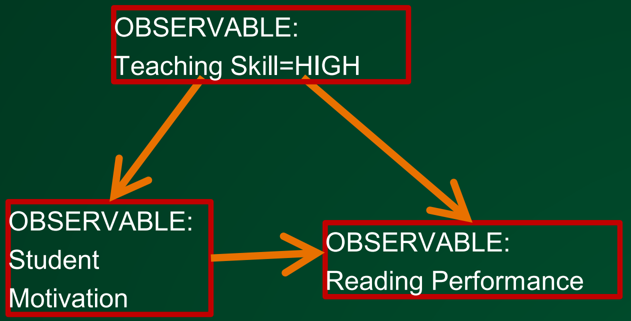
Confounder Problem in Causal Inference

Student Motivation	Reading Performance	Teaching Skill
3	8	3
0	4	0
1	2	0
2	3	5
1	4	1



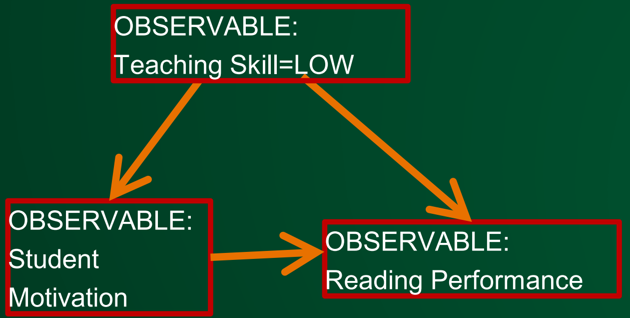
One Approach to Solving Confounder Problem in Causal Inference

Student Motivation	**Reading Performance**	**Teaching Skill**
3	8	3
2	3	5



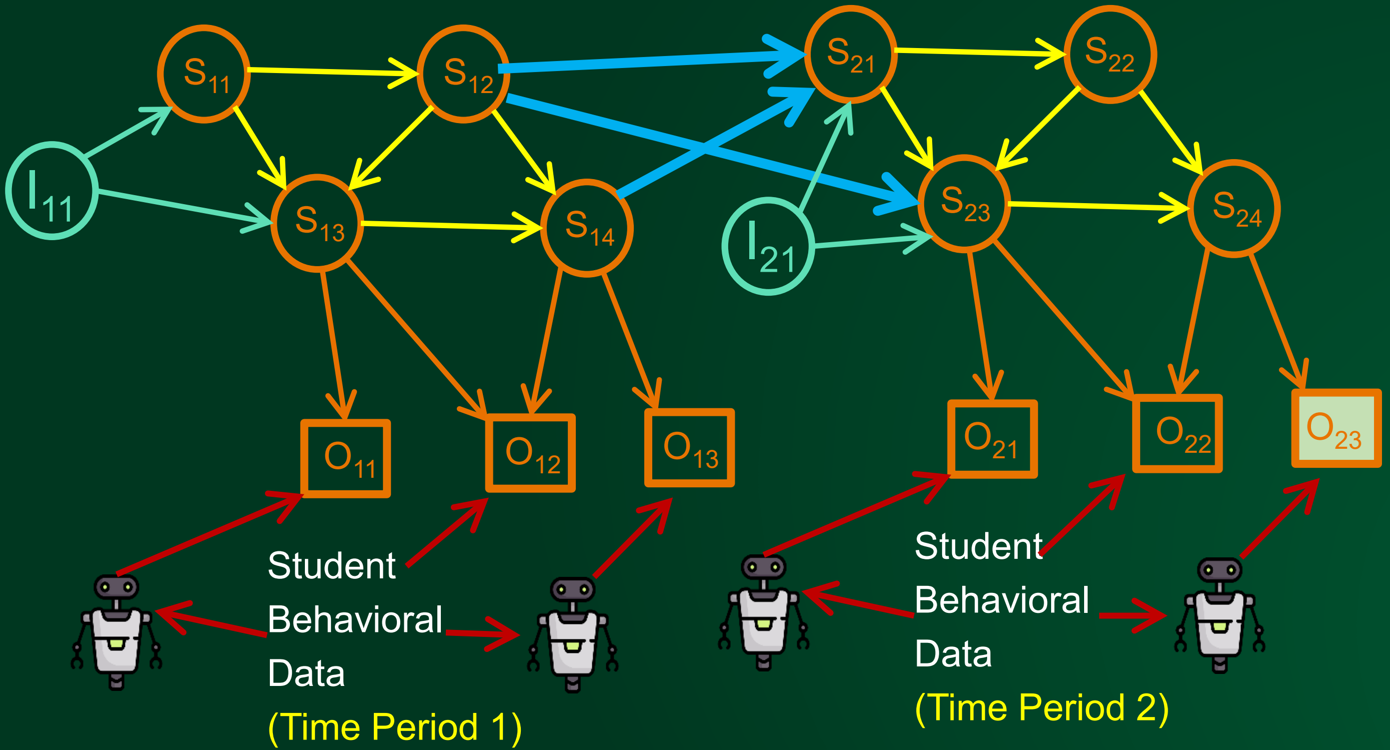
Effect of Student Motivation on Reading Performance Is Causally Identifiable

Student Motivation	**Reading Performance**	**Teaching Skill**
0	4	0
1	2	0
1	4	1

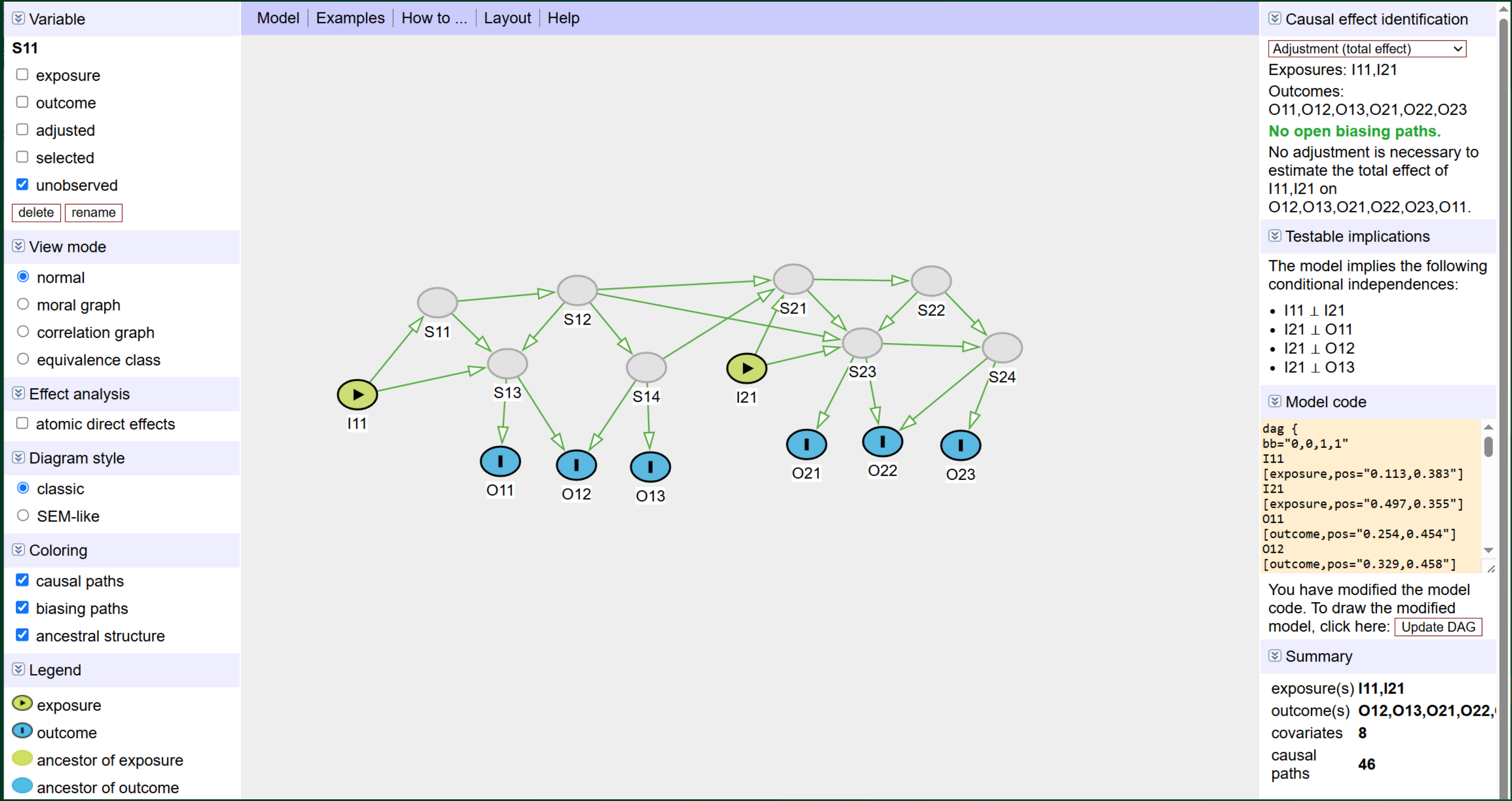


Cognitive Diagnostic Models: Causal Models

Does the Time 1 Intervention I1 **CAUSE** Student Behavioral Time 2 Outcome O3?



WEBSITE:https://www.dagitty.net/



WEBSITE:https://www.dagitty.net/

Variable

S11

- [x] exposure
- [ ] outcome
- [ ] adjusted
- [ ] selected
- [ ] unobserved

delete rename

View mode

- (•) normal
- ( ) moral graph
- ( ) correlation graph
- ( ) equivalence class

Effect analysis

- [ ] atomic direct effects

Diagram style

- (•) classic
- ( ) SEM-like

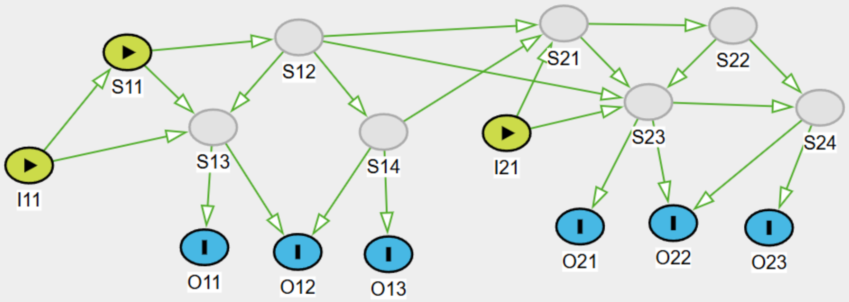
Coloring

- [x] causal paths
- [x] biasing paths
- [x] ancestral structure

Legend

- exposure
- outcome
- ancestor of exposure
- ancestor of outcome
- ancestor of exposure and

Model | Examples | How to ... | Layout | Help



Counterfactual Analysis:

Suppose everybody at time 1 has Skill 1?

Is the effect of the intervention causally identifiable?

What is the magnitude of the intervention effect?

Causal effect identification

Adjustment (total effect)

Exposures: I11,I21,S11

Outcomes: O11,O12,O13,O21,O22,O23

No open biasing paths.

No adjustment is necessary to estimate the total effect of I11,I21,S11 on O12,O13,O21,O22,O23,O11.

Testable implications

The model implies the following conditional independences:

- I11 ⊥ I21
- I11 ⊥ O13 | S11
- I11 ⊥ O21 | S11
- I11 ⊥ O22 | S11
- I11 ⊥ O23 | S11
- I21 ⊥ O11
- I21 ⊥ O12
- I21 ⊥ O13
- I21 ⊥ S11

Model code

```
S21 -> S22
S21 -> S23
S22 -> S23
S22 -> S24
S23 -> O21
S23 -> O22
S23 -> S24
S24 -> O22
S24 -> O23
}
```

Summary

exposure(s)	**I11,I21,S11**
outcome(s)	**O12,O13,O21,O22,**
covariates	**7**
causal paths	**76**

Summary

- Cognitive Diagnostic Models (CDMS) are AI Bayesian Networks
- Student Responses → AI→Evidence Indicators (observables in CDM)
- More complicated proficiency (e.g., Deep Proficiency models) and more complicated evidence models should continue to be explored as extensions of usual CDM. AI methods for inference and learning can be used.
- Dynamic Bayes Network Architecture is a CDM extension and well-studied in AI as a “planning algorithm” which is learned using “reinforcement learning”
- The strategy of reducing parameter dimensionality using DINO and DINA methods has been explored in AI in the areas “Markov Logic Nets” and “Statistical Relational AI”
- Continued work in statistical inference involving possibly misspecified, empirically non-identifiable, parameter redundant models is required as our psychometric models become more complicated…
- An important goal of assessment is to create a causal evidentiary reasoning chain and use that chain to guide curriculum design, automated AI personalized learning. Results from AI in the area of causal modeling are relevant here…

REFERENCE LIST FOR: Golden, R. M. (2025, April 26). *AI and Cognitive Diagnostic Models for Educational Assessment: Connections, Integrations, and New Directions*. [NCME organized session]. Bridging theory and Practice: Fostering innovation in diagnostic measurement to enhance education. National Council on Measurement in Education, Denver Colorado.

Almond, R. G. (2007). Cognitive modeling to represent growth (learning) using Markov decision processes. *Technology, Instruction, Cognition and Learning (TICL)*, *5*, 313-324. https://www.oldcitypublishing.com/TICL/TICL.html

Almond, R. G., DiBello, L., Jenkins, F., Senturk, D., Mislevy, R. J., Steinberg, L. S., & Yan, D. (2001). *Models for Conditional Probability Tables in Educational Assessment* Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research. https://proceedings.mlr.press/r3/almond01a.html

Almond, R. G., Robert J. Mislevy, Linda S. Steinberg, Duanli Yan, & Williamson, D. M. (2015). *Bayesian Networks in Educational Assessment*. Springer. https://link.springer.com/book/10.1007/978-1-4939-2125-6

Deonovic, B., Chopade, P., Yudelson, M., de la Torre, J., & von Davier, A. A. (2019). Application of Cognitive Diagnostic Models to Learning and Assessment Systems. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (pp. 437-460). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_21

Domingos, P., & Lowd, D. (2009). *Markov Logic: An interface layer for artificial intelligence*. Springer. https://link.springer.com/book/10.1007/978-3-031-01549-6

Flor, M., & Hao, J. (2021). Text Mining and Automated Scoring. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python* (pp. 245-262). Springer International Publishing. https://doi.org/10.1007/978-3-030-74394-9_14

Golden, R. M. (2020). *Statistical Machine Learning (Texts in Statistical Sciences Series)*. Chapman-Hall, CRC Press. https://www.routledge.com/Statistical-Machine-Learning-A-Unified-Framework/Golden/p/book/9781138484696

Golden, R. M. (2024). MAP Estimation Using a Possibly Misspecified Parameter Redundant Model. In M. Wilberg, J. Kim, H. Hwang, H. Wu, & T. Sweet (Eds.), *Quantitative Psychology: The 88th Annual Meeting of the Psychometric Society* (pp. 325-334). Springer Proceedings in Mathematics & Statistics. https://link.springer.com/chapter/10.1007/978-3-031-55548-0_30

Golden, R. M. (submitted). *Maximum likelihood estimation for possibly misspecified parameter redundant models*. https://labs.utdallas.edu/coinslab/files/2024/08/PsychoRedundancyV13_arxivstatTH-1.pdf

Green, D., Walsh, T., Cohen, P., & Chang, Y.-H. (2011). Learning a Skill-Teaching Curriculum with Dynamic Bayes Nets. *Proceedings of the AAAI Conference on Artificial Intelligence*, *25*, 1648-1654. https://doi.org/10.1609/aaai.v25i2.18855

Gu, Y. (2024). Going Deep in Diagnostic Modeling: Deep Cognitive Diagnostic Models (DeepCDMs). *Psychometrika*, *89*(1), 118-150. https://doi.org/10.1007/s11336-023-09941-6

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing*, *16*(2), 142-163. https://doi.org/10.1080/15305058.2015.1108322

Lauc, D., Grgic-Hlaca, N., & Skansi, S. (2016). Automatic Item Generation for Elementary Logic Quizzes via Markov Logic Networks. In (pp. 177-186). https://doi.org/10.4018/978-1-4666-9743-0.ch011

Levy, R. (2019). Dynamic Bayesian Network Modeling of Game-Based Diagnostic Assessments. *Multivariate Behavioral Research*, *54*(6), 771-794. https://doi.org/10.1080/00273171.2019.1590794

Madison, M. J., & Bradshaw, L. P. (2018). Assessing Growth in a Diagnostic Classification Model

Framework. *Psychometrika*, *83*(4), 963-990.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, *23*(2), 13-23. https://doi.org/10.3102/0013189X023002013

Mislevy, R., Behrens, J., DiCerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining*, *4*, 11-48.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*(4), 439-483. https://doi.org/10.1007/BF02294388

Pearl, J., Glymour, M., & Jewell, N. (2016). *Causal Inference in Statistics: A Primer*.

Raedt, L., Kersting, K., Natarajan, S., & Poole, D. (2016). *Statistical Relational Artificial Intelligence: Logic Probability, and Computation*. Springer. https://link.springer.com/book/10.1007/978-3-031-01574-8

Reichenberg, R. (2018). Dynamic Bayesian Networks in Educational Measurement: Reviewing and Advancing the State of the Field. *Applied Measurement in Education*, *31*(4), 335-350. https://doi.org/10.1080/08957347.2018.1495217

Song, Y., Junlei, D., & and Zheng, Q. Automatic item generation for educational assessments: a systematic literature review. *Interactive Learning Environments*, 1-20. https://doi.org/10.1080/10494820.2025.2482588

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. (2017). Robust causal inference using directed acyclic graphs: the R package ‘dagitty’. *International Journal of Epidemiology*, *45*(6), 1887-1894. https://doi.org/10.1093/ije/dyw341

Thompson, W. J., & Nash, B. (2022). A Diagnostic Framework for the Empirical Evaluation of Learning Maps [Original Research]. *Frontiers in Education*, *Volume 6 - 2021*. https://doi.org/10.3389/feduc.2021.714736

Wang, F., Gao, W., Liu, Q., Li, J., Zhao, G., Zhang, Z., Huang, Z., Zhu, M., Wang, S., Tong, W., & Chen, E. (2024). A survey of models for cognitive diagnosis: New developments and future directions. In. arxiv.

Weidlich, J., Gašević, D., & Drachsler, H. (2022). Causal Inference and Bias in Learning Analytics: A Primer on Pitfalls Using Directed Acyclic Graphs. *Journal of Learning Analytics*, *9*(3), 183-199. https://doi.org/10.18608/jla.2022.7577

Williamson, J. (2023). *Cognitive Diagnostic models and how they can be useful.* Cambridge University Press & Assessment. https://doi.org/10.17863/CAM.110853