Disentangling Person-Dependent and Item-Dependent Causal Effects

Josh Gilbert, Luke Miratrix, Mridul Joshi, Ben Domingue

2024-06-26

IRT HTE (Gilbert, 2024, JEBS)

1 Or: Every Interaction Effect You've Ever Estimated is Biased!*

*terms and conditions apply

IRT HTE (Gilbert, 2024, JEBS)

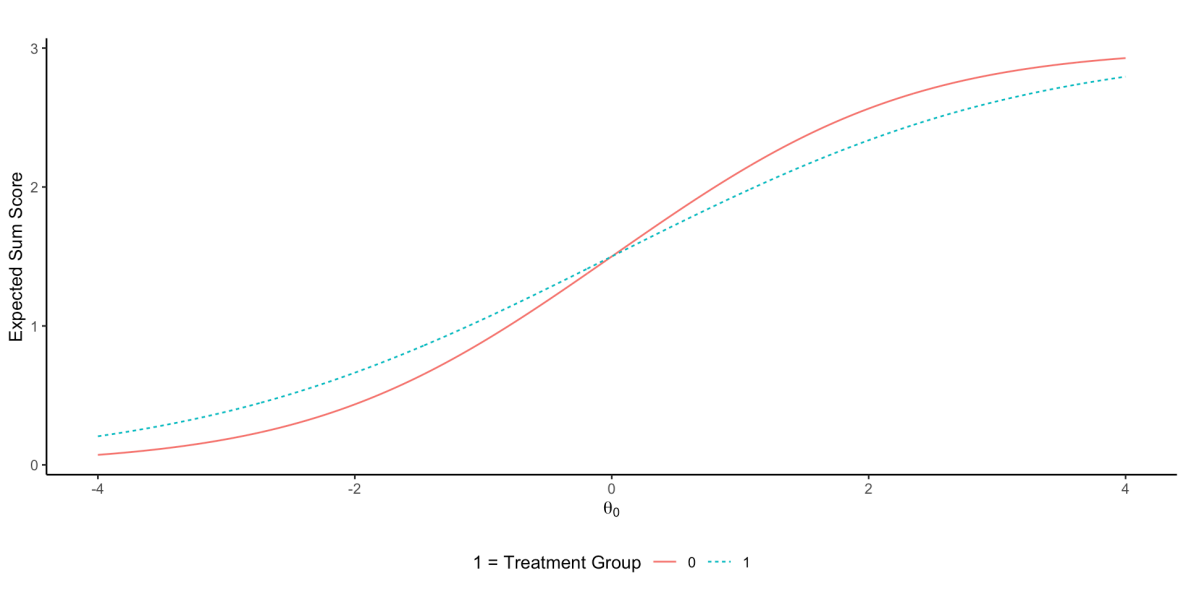
1.1 Prelude

- We have 60 minutes, it takes me about 35 to get through the slides uninterrupted
- Happy to take clarifying questions throughout; I may defer substantive questions until the end

2 Introduction

IRT HTE (Gilbert, 2024, JEBS)

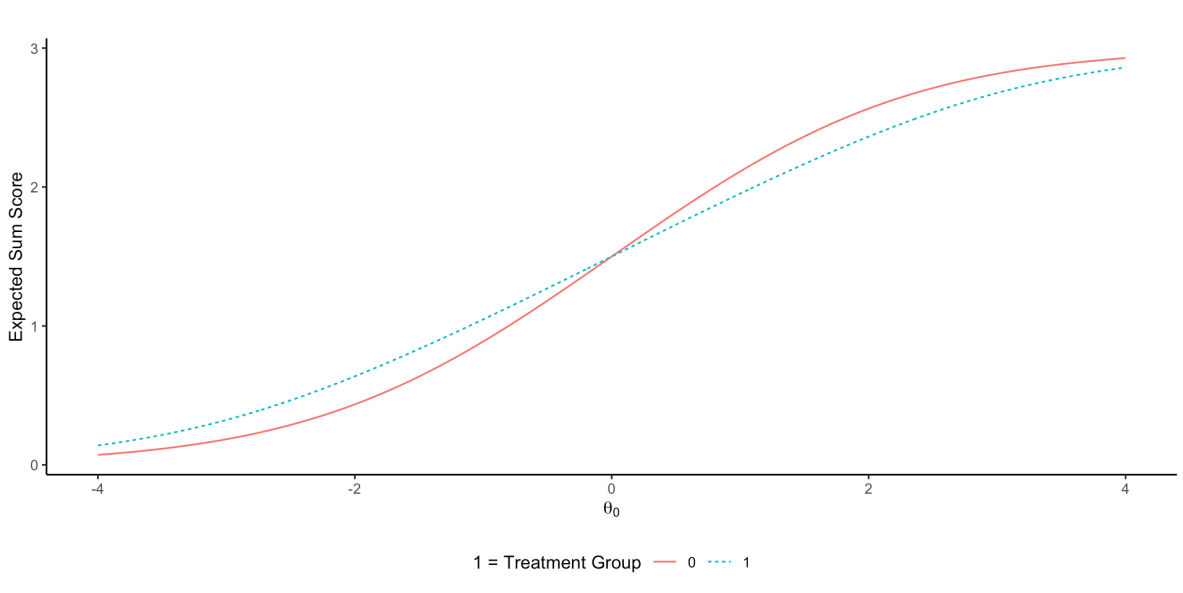
2.1 Let's look at some test score data from a (simulated) RCT!



- How would you interpret this?
- Most would say we have **heterogeneous treatment effects** (HTE) by pretest, or that there is a treatment by pretest **interaction**.
- Here, that's correct, the DGP here interacts treatment with pretest.

IRT HTE (Gilbert, 2024, JEBS)

2.2 How about this? (Not a trick question!)



- It looks the same right?
- However, the data were **not** generated with a treatment by baseline interaction!
- Here, the treatment effect size varies by item and is correlated with the item's easiness parameter (we will unpack what this means!)

IRT HTE (Gilbert, 2024, JEBS)

2.3 What is going on here?

- There are two sources of HTE, **person** dependent (Graph 1) and **item** dependent (Graph 2), that are **empirically indistinguishable** with standard methods when analyzing test scores
- We really (should) care about this, because understanding HTE allows us to target interventions towards populations in which they'd be most effective, which is critical for **policy relevance**

IRT HTE (Gilbert, 2024, JEBS)

3 A Tale of Two Data-Generating Processes (DGPs)

IRT HTE (Gilbert, 2024, JEBS)

3.1 But first, a Model of Treatment Effects (in IRT)

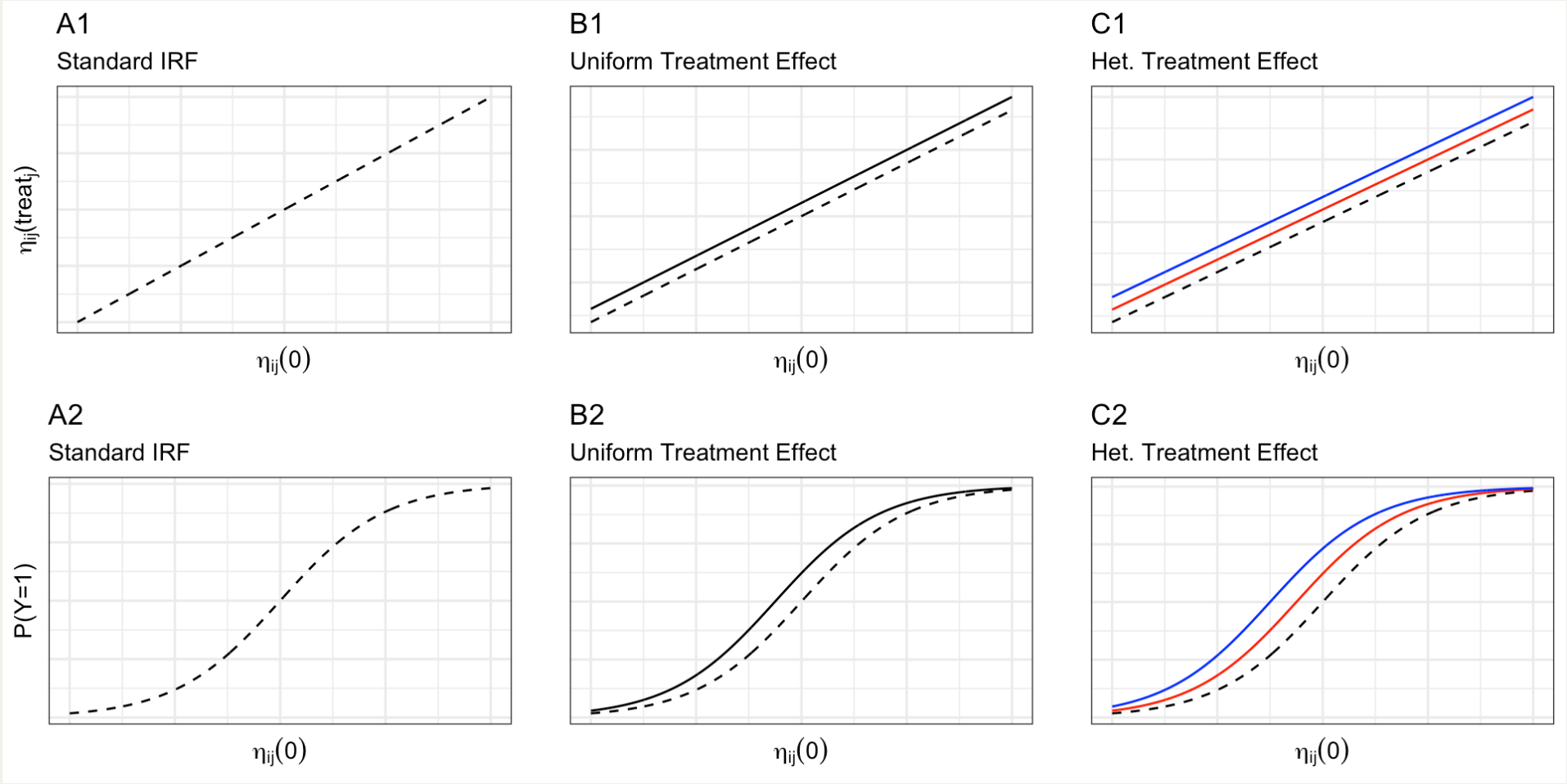
Consider a standard Rasch IRT model

$$\Pr(y_{ij} = 1) = \text{logit}^{-1}(\eta_{ij}) = \text{logit}^{-1}(\theta_j^1 + b_{ij})$$

We can define our treatment effect τ_{ij} as the difference in potential outcomes in terms of the linear predictor η_{ij}:

$$\tau_{ij} = \eta_{ij}(1) - \eta_{ij}(0)$$

3.2 Intuition: Item Response Functions and τ_{ij}



IRT HTE (Gilbert, 2024, JEBS)

3.3 Two DGPs, Two Types of HTE

θ_j^0 is baseline ability

Person-Dependent HTE

$$\theta_j^1 = \beta_0 + \beta_1 \text{treat}_j + \beta_2 \theta_j^0 + \beta_3 \text{treat}_j \times \theta_j^0$$
$$b_i \sim N(0, \sigma_0)$$

(e.g., remedial instruction targeted at poor performers)

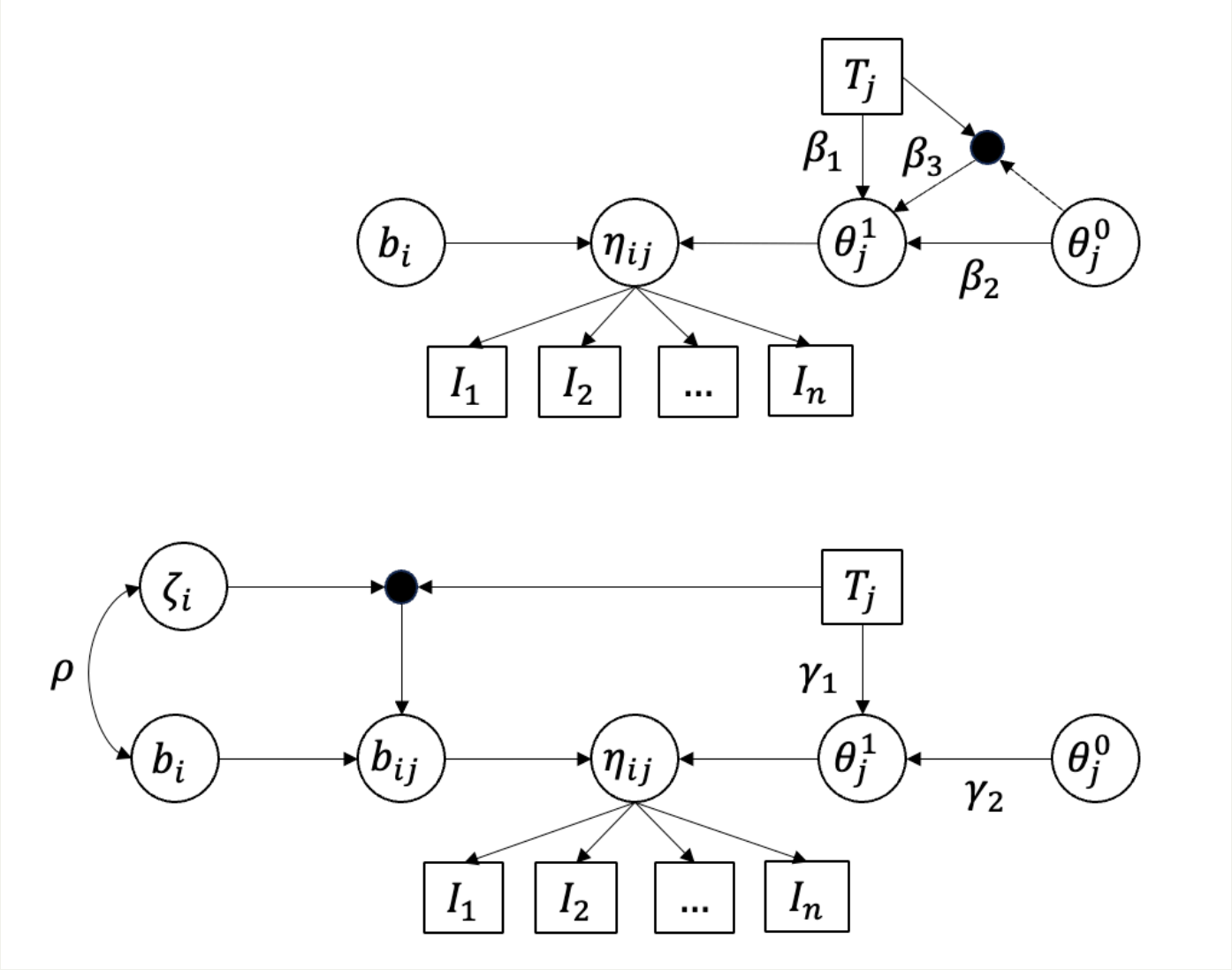
Item-Dependent HTE

$$\theta_j^1 = \gamma_0 + \gamma_1 \text{treat}_j + \gamma_2 \theta_j^0$$
$$b_{ij} = b_i + \zeta_i \text{treat}_j$$
$$\begin{bmatrix} b_i \\ \zeta_i \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_0 & \varrho \\ \varrho & \sigma_1 \end{bmatrix}\right)$$

(e.g., basic skills instruction)

Essentially, our argument is that β_3 in the person model and ϱ in the item model can become **confounded** (i.e., unidentified) and can lead to **identical** patterns of observed HTE at the test score level.

3.4 DAG Representation



IRT HTE (Gilbert, 2024, JEBS)

4 Confounding of β_3 and ϱ

Several convergent lines of reasoning:

1. Toy Example
2. Visual
3. Analytic
4. Computational

IRT HTE (Gilbert, 2024, JEBS)

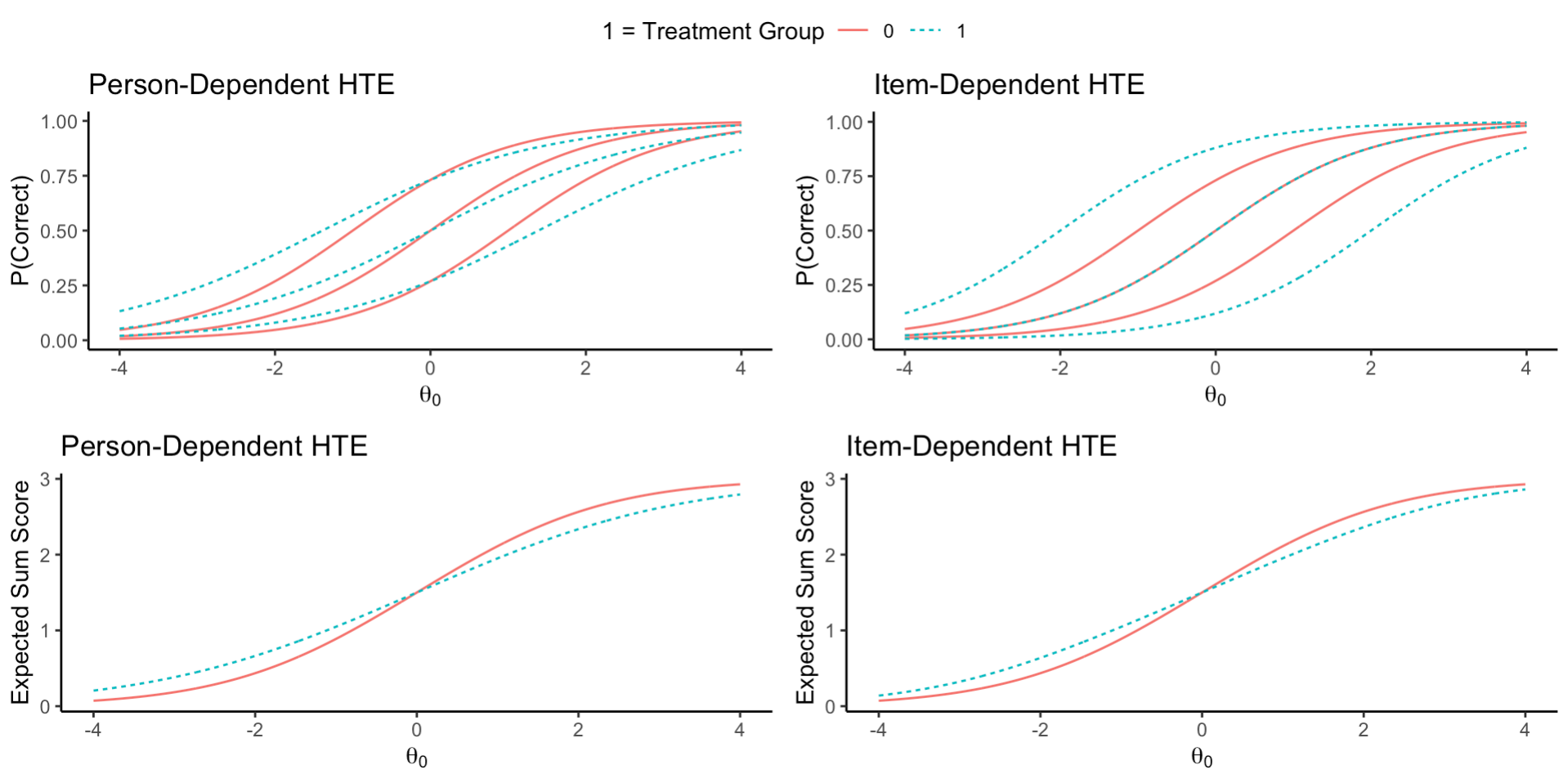
4.1 A Toy Example

- Set $\beta_0 = 0, \beta_1 = 0, \beta_2 = 1, \gamma_1 = 0, \gamma_2 = 1$. Person model, set $\beta_3 = -.28$, Item model, set $\varrho = 1$ so that $\zeta_i = b_i$.
- Three items, $b_i = -1, 0, 1$ and three people, $\theta_j^0 = -1, 0, 1$

Table 1: Toy example illustrating the identification problem on a three-item test

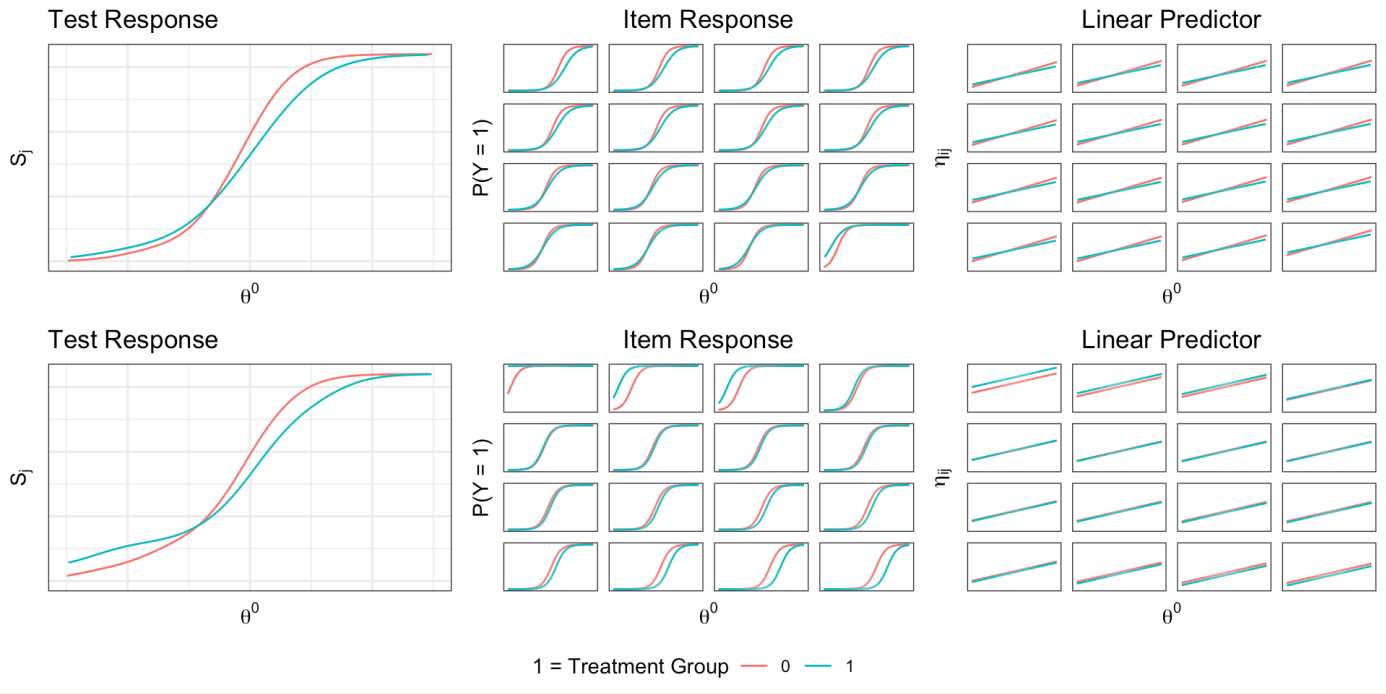
j	θ_j^0	$\mathbb{E}_P(S_j(0))$	$\mathbb{E}_P(S_j(1))$	$\mathbb{E}_I(S_j(0))$	$\mathbb{E}_I(S_j(1))$	τ_P	τ_I
1	-1	0.88	1.04	0.88	1.04	0.16	0.16
2	0	1.50	1.50	1.50	1.50	0.00	0.00
3	1	2.11	1.95	2.11	1.95	-0.16	-0.16

4.2 A Toy Example



IRT HTE (Gilbert, 2024, JEBS)

4.3 A Visual Proof



Some foreshadowing: while the patterns in the total test score are indistinguishable, the item-level patterns are starkly different, suggesting that an appropriate item-level analysis could **disentangle** the causes of the HTE.

IRT HTE (Gilbert, 2024, JEBS)

4.4 A Mathematical Proof

- Under some simplifying assumptions, we can show that the two DGPs can generate **identical** estimates of the treatment effect as a function of baseline ability. i.e., for any values of $\beta_1, \beta_2, \beta_3$ I set $\varrho = 1$ and provide values of $\gamma_0, \gamma_1, \gamma_2$ that give us the same τ_j (not scores themselves, but treatment effects)
- This is true whether we use a sum score, IRT-based score, or a (misspecified) latent variable model
- See paper for details
- (I later found a different approach that works by matching the test score curves themselves; this is at the end if time/interest)

4.5 The (Proposed) Solution

We can fit a flexible model that allows for **both** the treatment by baseline interaction β_3 **and** the treatment by item easiness correlation ϱ:

$$\theta_j^1 = \beta_0 + \beta_1 \text{treat}_j + \beta_2 \theta_j^0 + \beta_3 \text{treat} \times \theta_j^0$$

$$b_{ij} = b_i + \zeta_i \text{treat}_j$$

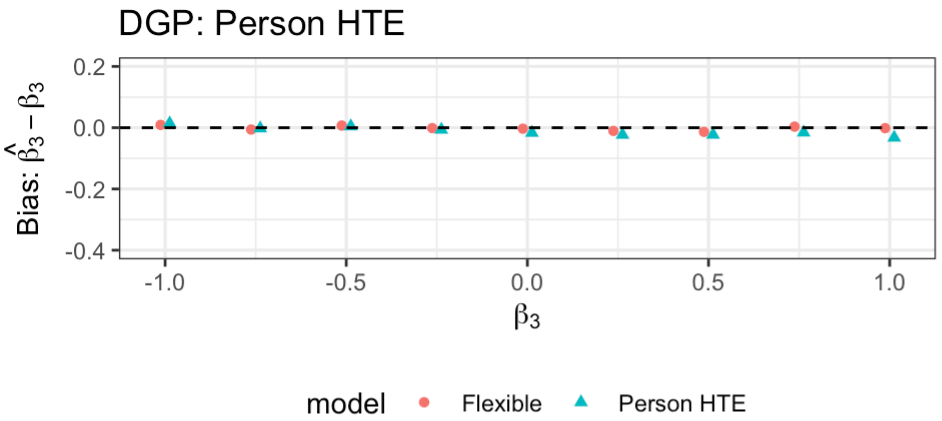
$$\begin{bmatrix} b_i \\ \zeta_i \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_0 & \varrho \\ \varrho & \sigma_1 \end{bmatrix}\right)$$

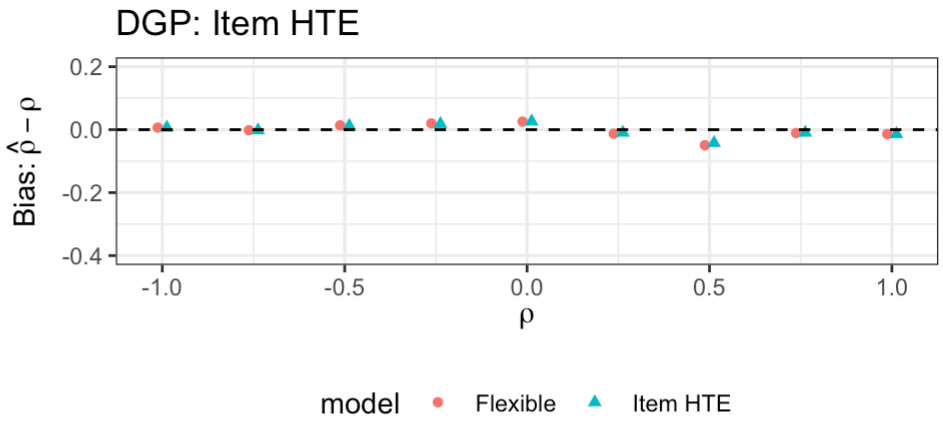
IRT HTE (Gilbert, 2024, JEBS)

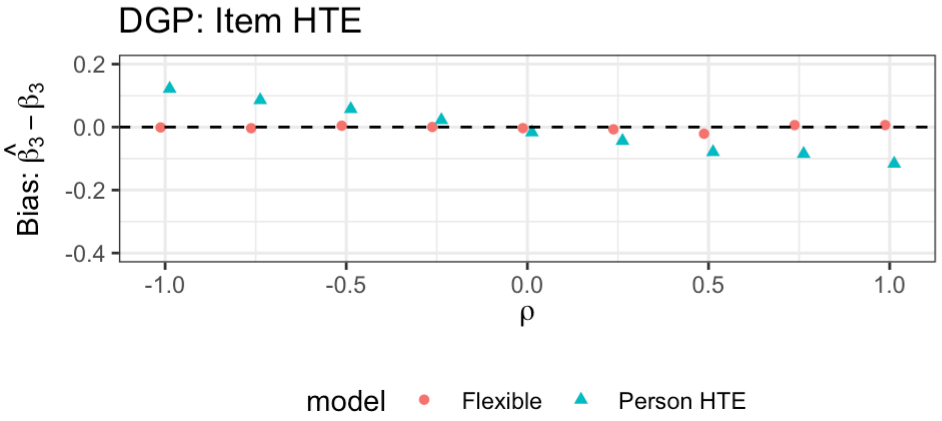
4.6 Monte Carlo Simulation

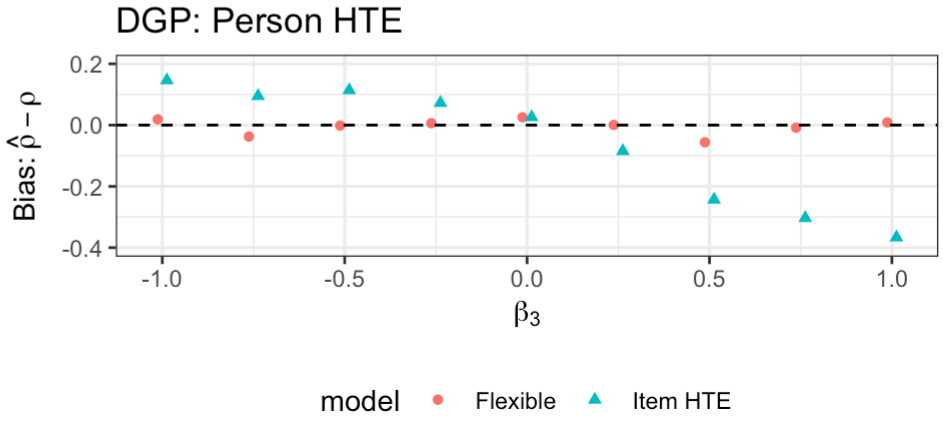
- We can test under more realistic conditions
- We will see how β_3 and ϱ become confounded
- We will see how the flexible model can determine the correct DGP

4.7 Simulation Results









IRT HTE (Gilbert, 2024, JEBS)

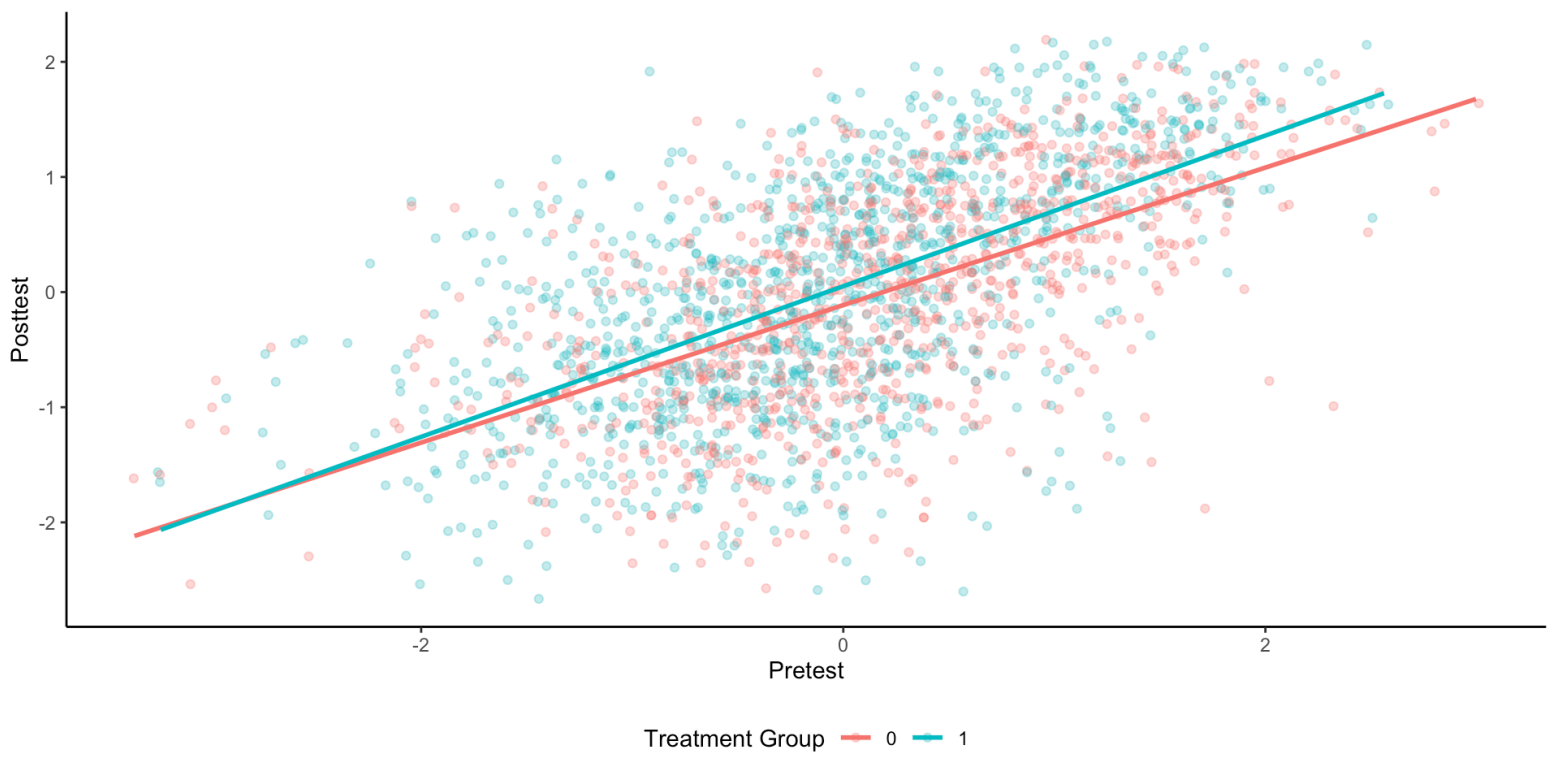
5 Empirical Application

IRT HTE (Gilbert, 2024, JEBS)

5.1 Model of Reading Engagment (MORE) Data

- Public use file that includes item-level data (Kim et al., 2023, JEP)
- MORE content literacy intervention in 1st and 2nd grade, N = 2174, cluster randomized trial
- Outcome: researcher-developed reading comprehension assessment with 20 dichotomous items
- Original paper showed positive average treatment effects (about +0.18 SDs) on reading comprehension
- Let's analyze for HTE!

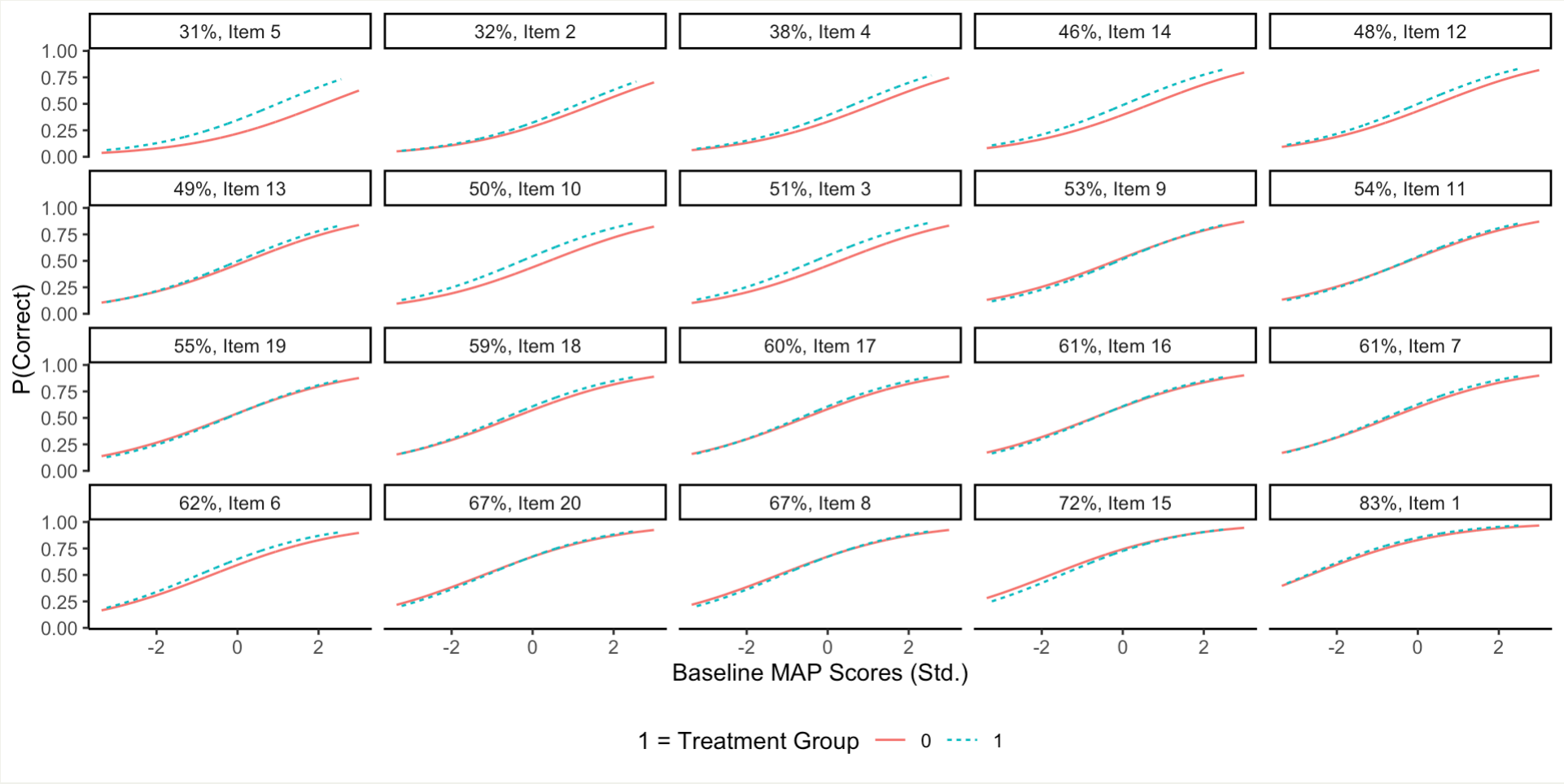
5.2 EDA (1) - What should we conclude?



- If you've been paying attention, nothing yet!

IRT HTE (Gilbert, 2024, JEBS)

5.3 EDA (2) - Item Level



IRT HTE (Gilbert, 2024, JEBS)

5.4 Selected Regression Models

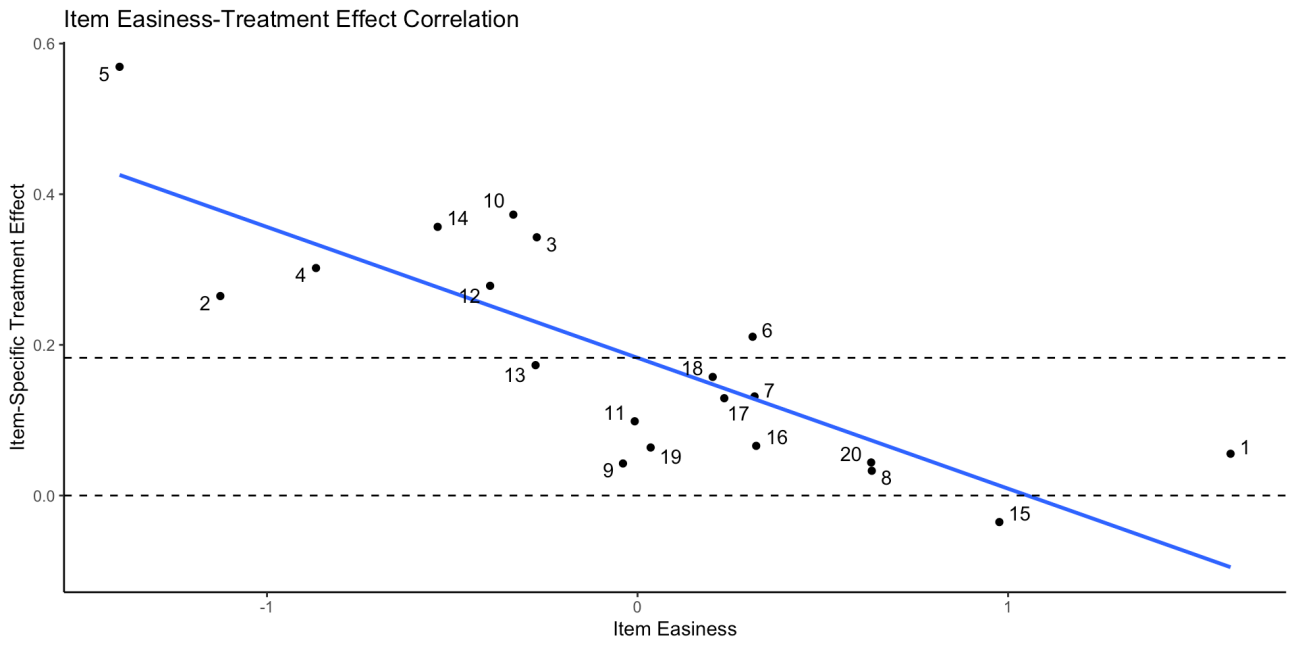
	Person HTE	Flexible
Constant	0.12 (0.15)	0.12 (0.16)
Treat	0.18 (0.04) ***	0.18 (0.05) ***
Std. Baseline	0.61 (0.03) ***	0.62 (0.03) ***
Treat x Baseline	0.08 (0.04) *	0.06 (0.04)
AIC	52343.10	52309.33
BIC	52395.18	52378.77
Log Likelihood	-26165.55	-26146.66
Num. obs.	43480	43480
Num. groups: s_id	2174	2174
Num. groups: s_q_num	20	20
Var: s_id (Intercept)	0.46	0.46
Var: s_q_num (Intercept)	0.41	0.50
Var: s_q_num s_itt_consented		0.03
Cov: s_q_num (Intercept) s_itt_consented		-0.09

*** p < 0.001; ** p < 0.01; * p < 0.05

IRT HTE (Gilbert, 2024, JEBS)

5.5 Correlation of Item Easiness and Treatment Effect Size

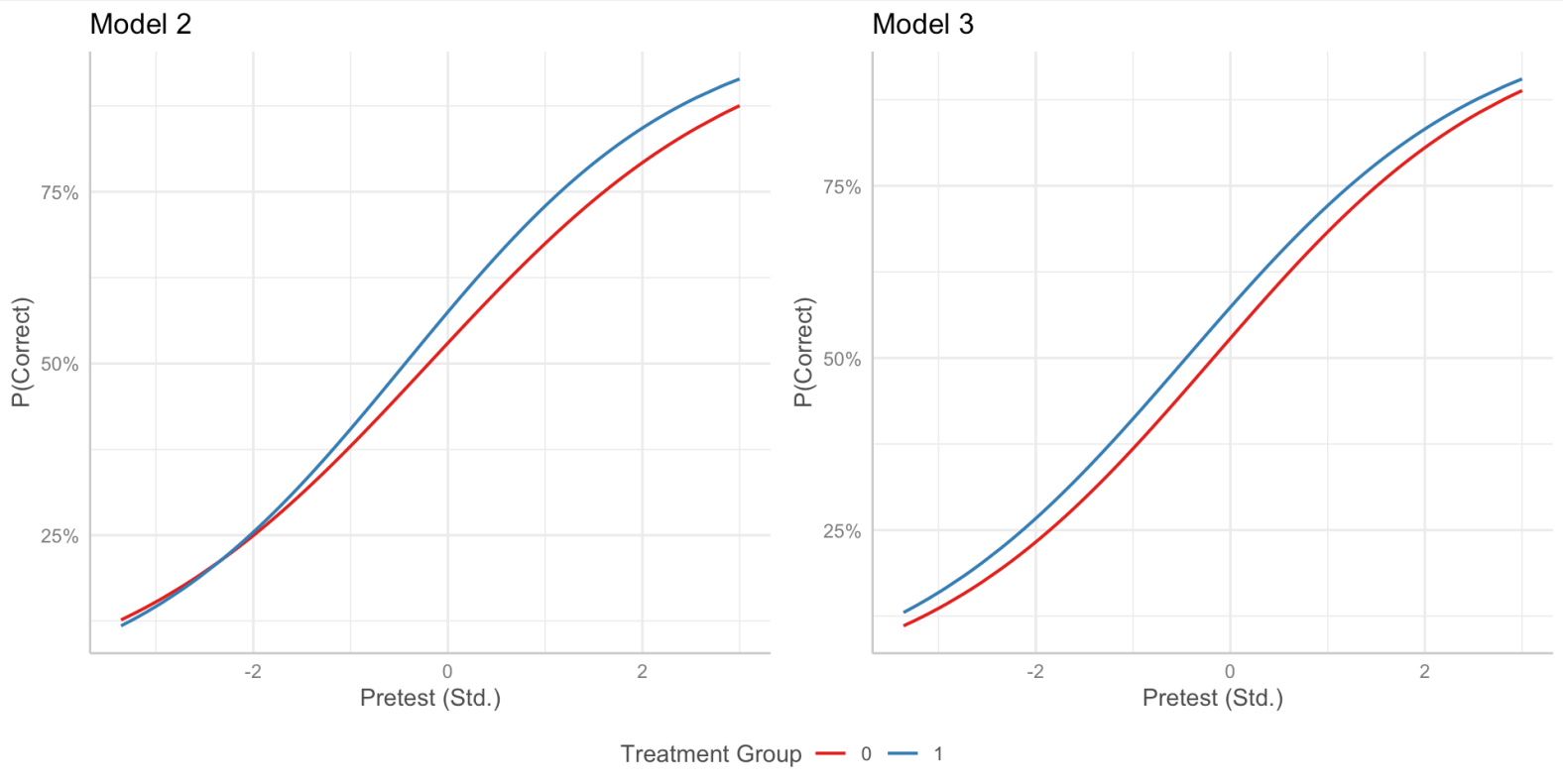
MORE had the largest effects on the **most difficult items** that better distinguish among higher ability students, $\hat{\varrho} = -0.74$



IRT HTE (Gilbert, 2024, JEBS)

5.6 Graphical Interpretation

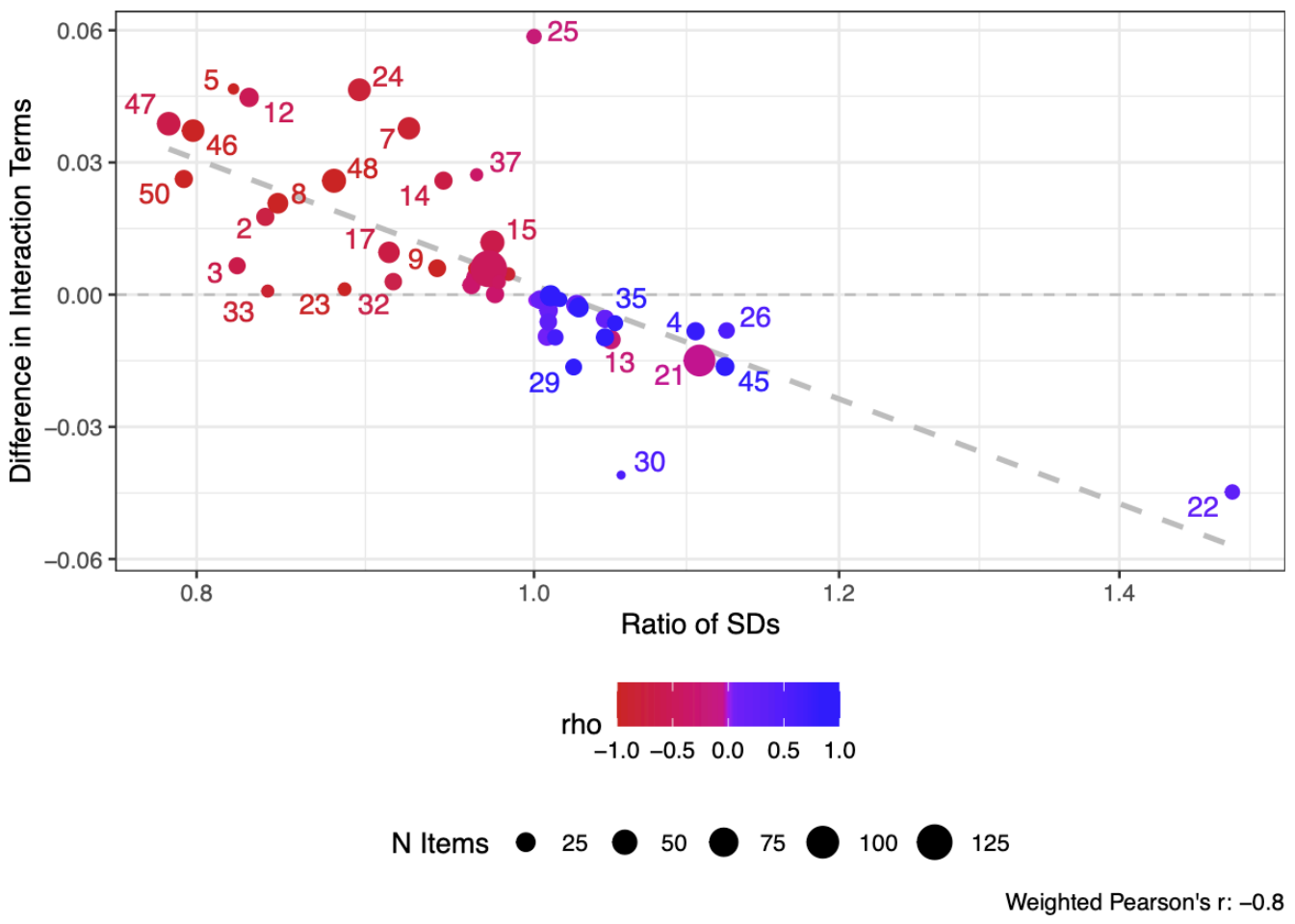
The traditional approach would point to the world on the left, the flexible approach to the world on the right. The policy implications are clearly different!



IRT HTE (Gilbert, 2024, JEBS)

5.7 How Prevalent is This?

50 item-level datasets from RCTs (collaboration opportunity!)



IRT HTE (Gilbert, 2024, JEBS)

6 Conclusions

A treatment that improves proficiency for low achieving students (across all items) has different implications than one that improves performance on easier content (across all students). Where total test scores obscure this distinction, item-level analysis can tell us what's really happening.

IRT HTE (Gilbert, 2024, JEBS)

6.1 Summary (1)

- **Most people** look at total test scores when evaluating HTE, but treatment effects on total test scores can be a product of very different mechanisms of treatment impact
- **Clever people** might use item level-data in a latent variable model, but most LVMs ignore item-dependent HTE
- **Extremely clever people** (like us, and perhaps, you) will fit a model that allows for both kinds of variation because the items allow for disambiguation

IRT HTE (Gilbert, 2024, JEBS)

6.2 Summary (2)

- This is not just a theoretical concern, as the MORE data shows
- Researchers should share item-level data (IRW)
- The example of pretest scores used here is generalizable to **any** covariate that is correlated with the outcome (e.g., age, sex, SES), **any** treatment variable, and **polytomous** responses, so our method is applicable to many fields beyond education (economics, psychology, health, etc.)
- **IRT provides a powerful (and mostly unexplored) tool for causal analysis, not just test design**

IRT HTE (Gilbert, 2024, JEBS)

6.3 Thank you!

josh.b.gilbert@gmail.com / joshua_gilbert@g.harvard.edu

JEBS Paper

Replication Toolkit

(Happy to share PDFs)

7 Appendix

Simplified Proof

7.1 A (Simpler) Mathematical Proof

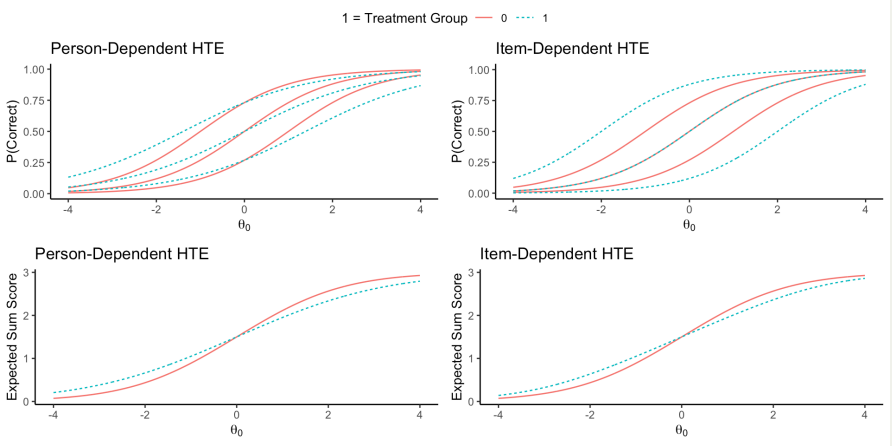
IRT HTE (Gilbert, 2024, JEBS)

7.2 A (Simpler) Mathematical Proof

$$\beta_{TRF}|(T_j = 1) = \frac{\beta_2 + \beta_3}{\sqrt{.346\sigma_0^2 + 1}}$$

$$\beta_{TRF}|(T_j = 1) = \frac{\gamma_2}{\sqrt{.346\sigma_0^{2*} + 1}}$$

$$\sigma_0^{2*} = \sigma_0^2 + \sigma_1^2 + 2\varrho\sigma_0\sigma_1$$



- So theoretically, we could see the same effect if $\varrho = 0$, but σ_1 would have to be enormously large
- ϱ acts as a "multiplier" that allows for the items to either "stretch" or "compress" the horizontal axis

- Assuming $\beta_2 = \gamma_2$, whatever value you choose for β_3, I can choose (infinitely) many sets of σ_1 and ϱ that would generate the identical pattern (to a point; σ_1 must be positive)
- (The proof in the paper follows this same logic but has a slightly different target)