Modeling Item-Level Heterogeneous Treatment Effects with the Explanatory Item Response Model

Josh Gilbert, Jimmy Kim, and Luke Miratrix (Harvard GSE)

JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

0.1 Materials

JEBS Publication

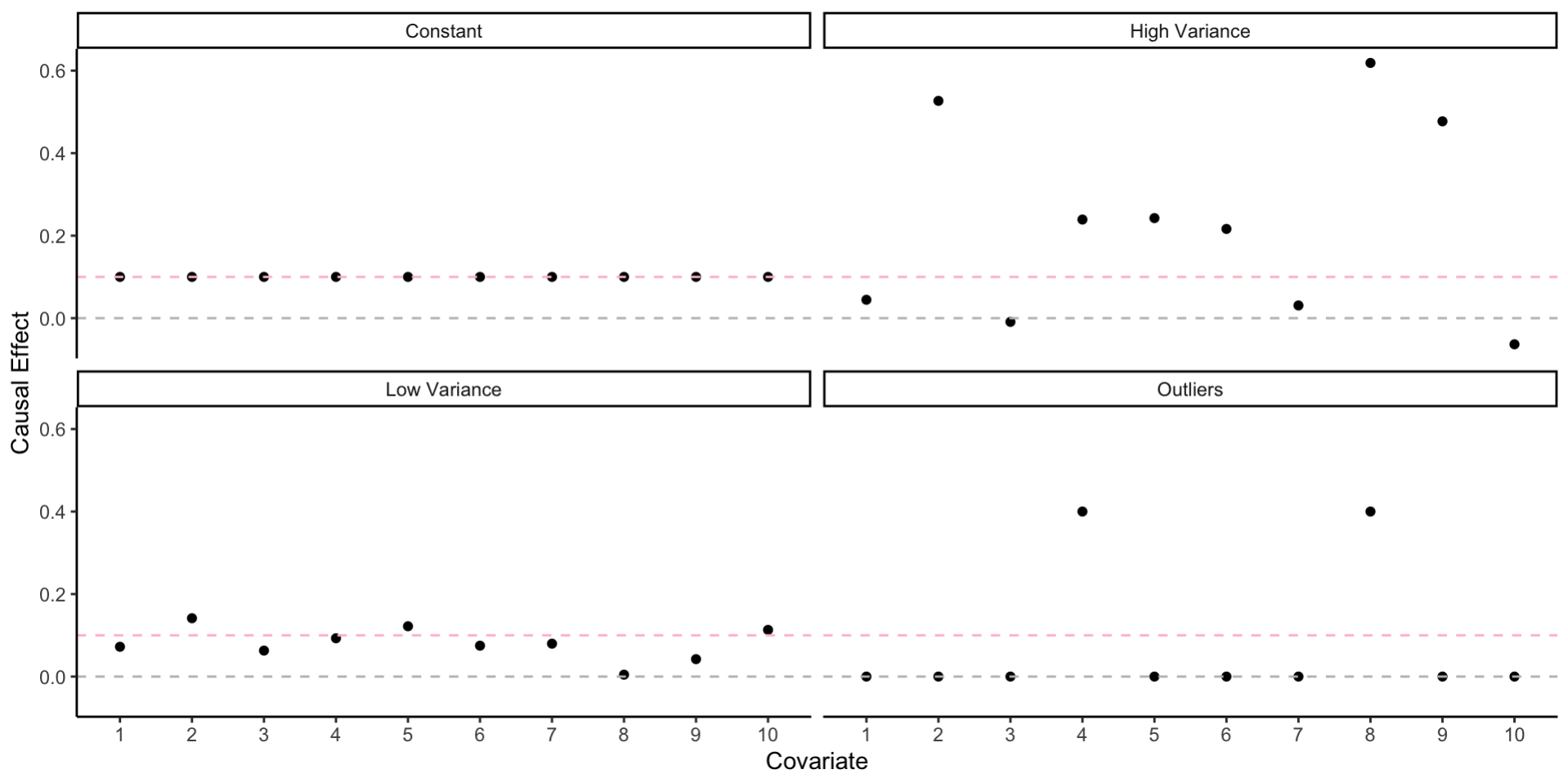
Replication Toolkit

1 Introduction

1.1 The Importance of Heterogeneous Treatment Effects (HTE)

How can we move beyond the traditional "what works" model of intention-to-treat analysis to address questions of "for whom" and "under what conditions" an intervention succeeds?

1.2 Gelman's (2022) Causal Quartet



JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

1.3 Existing Statistical Methods for Evaluating HTE

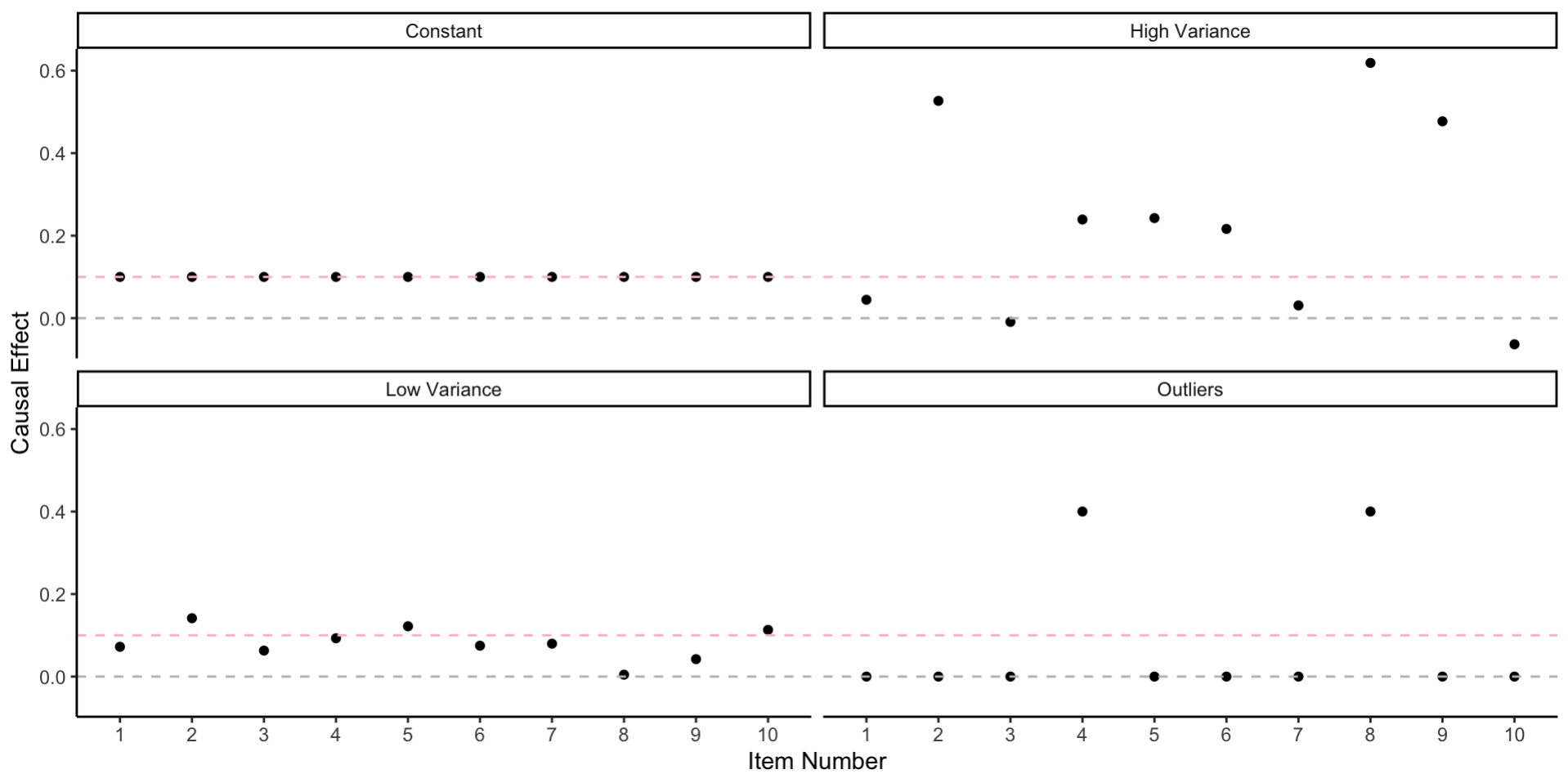
- Subgroup Analysis
- Statistical Interactions/Moderation
- Quantile Regression
- Instrumental Variables
- Mediation
- All of these techniques are (typically) focused on the **person** level

1.4 Might Treatment Differentially Impact Assessment Items?

- The contribution of this study is to propose and demonstrate the utility of a statistical method that allows for **within-outcome**, or "Item-Level" HTE (IL-HTE) using the **Explanatory Item Response Model** (EIRM).

- We demonstrate with a data simulation and application to empirical data from a cluster randomized trial.

1.5 Gelman's (2022) Causal Quartet



JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

2 Statistical Model

2.1 Intuition 1: *Person-Level* Data Structure

Consider the following data structure:

id	treat	score
1	0	50
2	0	55
3	1	60
4	1	65

```
lm(score ~ treat, data)
```

2.2 Intuition 2: *Item-Level* Data Structure

Consider this alternative data structure representing the same data. Underneath each score are the item responses, which we can **model directly**.

id	item	treat	correct
1	1	0	0
1	2	0	0
1	3	0	1
3	1	1	0
3	2	1	1
3	3	1	1

2.3 Intuition 2: *Item-Level* Data Structure

We can estimate the average treatment effect at the **item-level** with a **cross-classified logistic regression** model:

```
glmer(correct ~ treat + (1|item) +
(1|student), data, binomial)
```

Mathematically, this is equivalent to a 1PL IRT model with a treatment effect. These models have historically been used in descriptive, not causal, contexts (see De Boeck & Wilson, 2004).

2.4 Formalizing the Statistical Model

OLS Regression: $\text{score}_j = \beta_0 + \beta_1 \text{treat}_j + \varepsilon_j$

Explanatory Item Response Model (EIRM, cross-classified logistic regression):

$$\text{logit}(P(\text{correct}_{ij} = 1)) = \beta_0 + \beta_1 \text{treat}_j + \theta_j + \zeta_i$$

β_0 = log-odds of the correct response in the control group

β_1 = treatment effect in logits

θ = residual latent person ability, assumed $N(0, \sigma^2_\theta)$

ζ = residual latent item easiness, assumed $N(0, \sigma^2_\zeta)$

2.5 Allowing for Item-Level HTE (IL-HTE)

$$\text{logit}(P(\text{correct}_{ij} = 1)) = \beta_0 + \beta_1 \text{treat}_j + \theta_j + \zeta_{0i} + \zeta_{1i} \text{treat}_j$$

$$\theta \sim N(0, \sigma^2_\theta)$$

$$\begin{bmatrix} \zeta_0 \\ \zeta_1 \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma^2_{\zeta 0} & \varrho_{01} \\ \varrho_{10} & \sigma^2_{\zeta 1} \end{bmatrix})$$

- We add a **random slope** for treatment at the item level (ζ_{1i}), which can be considered a type of **Differential Item Functioning** (DIF) caused by the intervention
- We obtain the **average treatment effect** (β_1) and the distribution of item-specific treatment effects ($\sigma^2_{\zeta_1}$)

2.6 Allowing for Item-Level HTE (IL-HTE)

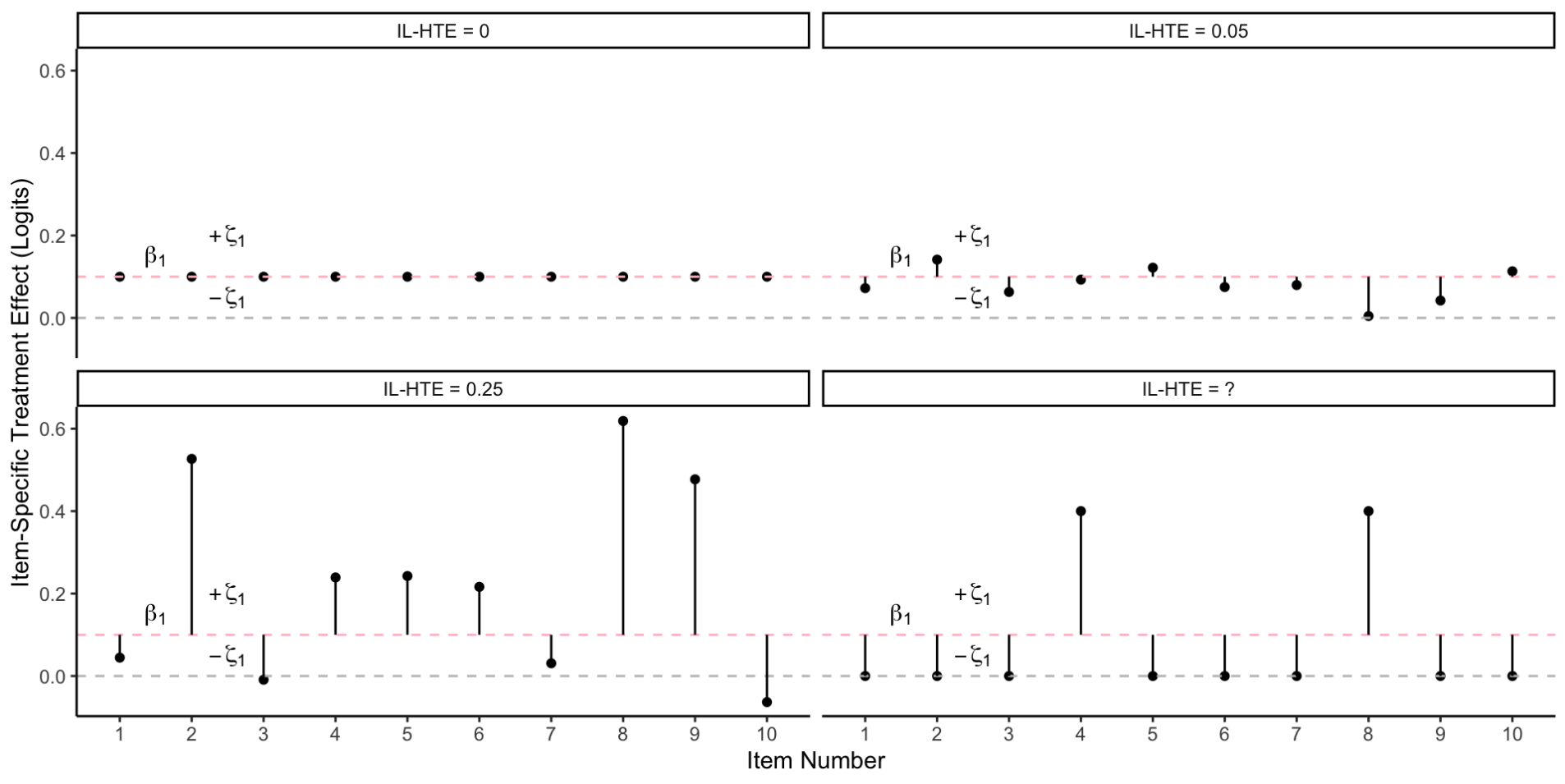
$$\text{logit}(P(\text{correct}_{ij} = 1)) = \beta_0 + \beta_1 \text{treat}_j + \theta_j + \zeta_{0i} + \zeta_{1i} \text{treat}_j$$

$$\theta \sim N(0, \sigma^2_\theta)$$

$$\begin{bmatrix} \zeta_0 \\ \zeta_1 \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma^2_{\zeta 0} & \varrho_{01} \\ \varrho_{10} & \sigma^2_{\zeta 1} \end{bmatrix})$$

- We can add treatment by item characteristic interactions to capture **systematic IL-HTE**; $\sigma^2_{\zeta_1}$ represents **unexplained IL-HTE**.
- For example, certain **item-clusters** or **subscales** may show greater responsiveness to an intervention, which could be captured by the interaction effect (even if the scale is unidimensional)

2.7 Allowing for Item-Level HTE (IL-HTE)



JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

3 Methods

Monte Carlo Simulation and Empirical Application

3.1 Monte Carlo Simulation

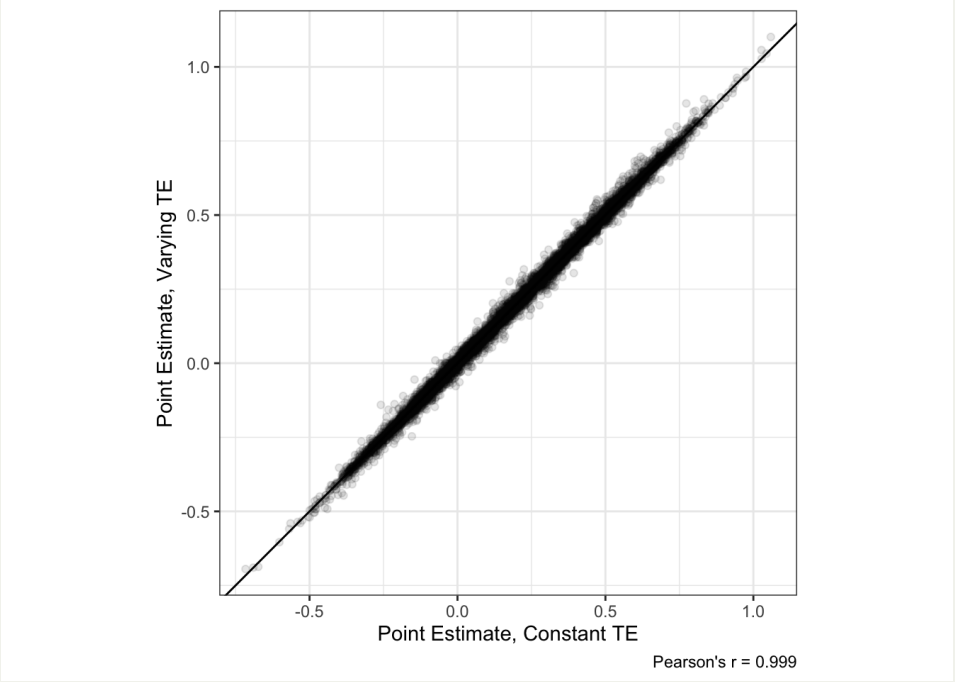
- We used Monte Carlo simulation to test the performance of the EIRM to model IL-HTE compared to a model that assumes a constant treatment effect:
  - Bias for $\hat{\beta_1}$
  - Calibration of $SE(\hat{\beta_1})$
  - False positive rates for $\hat{\beta_1}$

3.2 Simulation Design

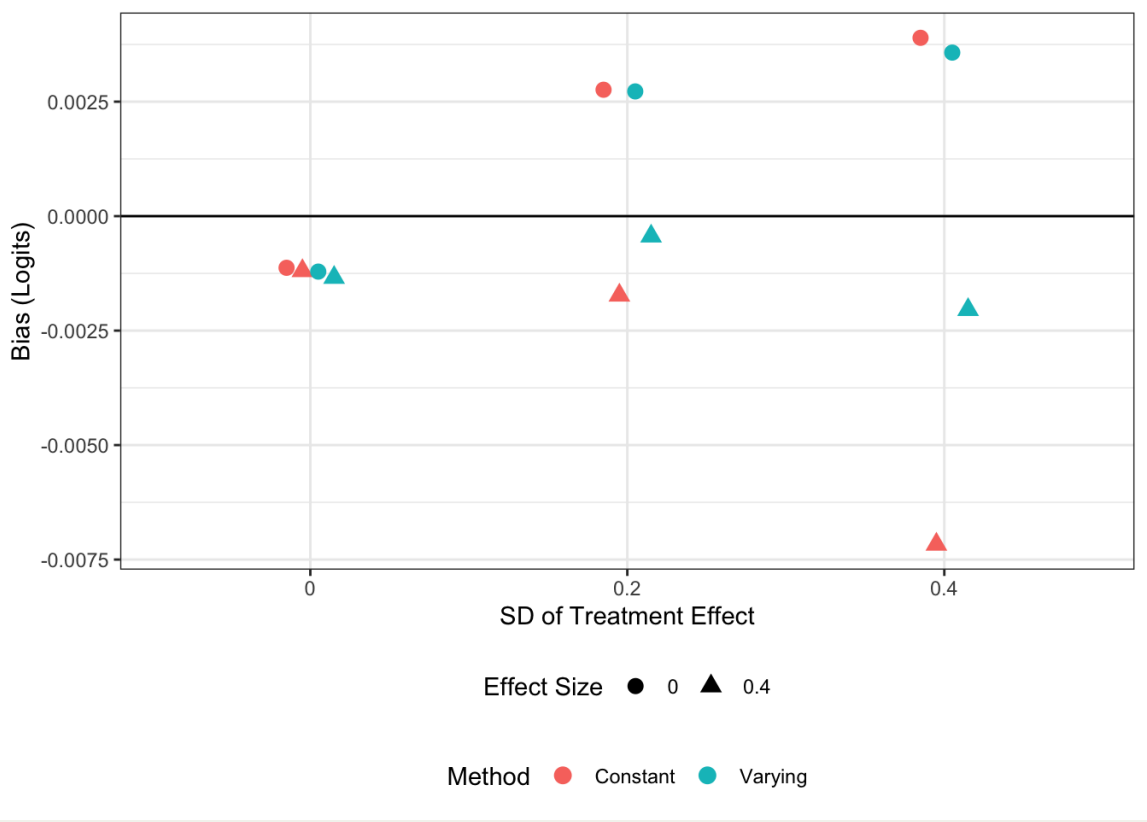
- **Fixed Factors**: 500 subjects with ability θ_j fully crossed with 20 items with easiness ζ_{0i}, both drawn from $N(0, 1)$
- **Varying Factors:** Null (0) and positive (0.4) ATEs β_1 fully crossed with no (0), moderate (0.2), and high (0.4) IL-HTE σ_{ζ_1}
- Item responses generated from a Rasch model with a treatment effect: $P(correct_{ij}) = logit^{-1}[(\theta_j + \zeta_{0i}) + (\beta_1 + \zeta_{1i})treat_j]$
- 2000 replicates per condition, 12,000 data sets total
- Analyzed with both random intercept (constant effect) and random slopes (IL-HTE) models

4 Simulation Results

4.1 Simulation Result 1: Point Estimates for $\hat{\beta_1}$ are Essentially Identical Across Models



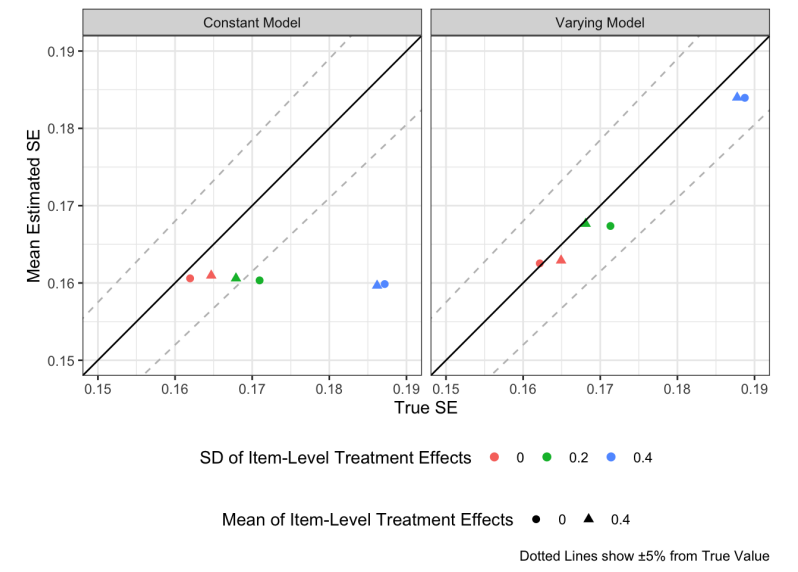
4.2 Simulation Result 2: Bias for $\hat{\beta_1}$ is Negligible Across Models



JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

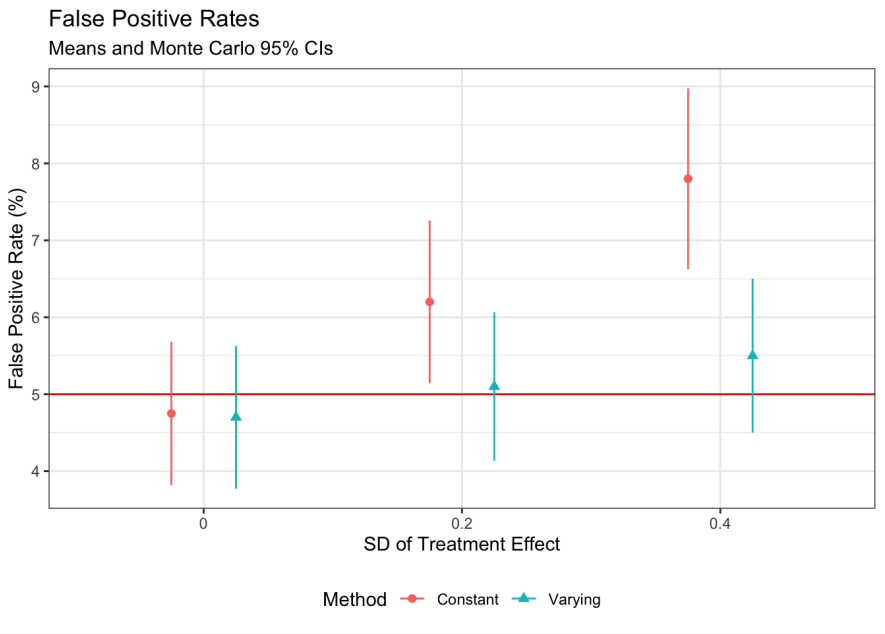
4.3 Simulation Result 3: $SE(\hat{\beta_1})$ are Underestimated when IL-HTE ($\sigma^2_{\zeta_1}$) is Ignored



4.4 Simulation Result 3: $SE(\hat{\beta_1})$ are Underestimated when IL-HTE ($\sigma^2_{\zeta_1}$) is Ignored

- SEs from the constant effect model fail to take into account the selection of items onto a test from a broader (real or hypothetical) **pool of items**.
- The constant effect model is targeting the ATE for the **specific items** that happen to be on the assessment, not the **population of items** from which a test was constructed.
- If IL-HTE is 0, this doesn't matter, but if it is greater than 0, **SEs are underestimated**
- This builds on previous findings about random slopes in mixed models generally, or fixed vs. random effects estimators in meta-analysis or multi-site trials (see Miratrix, et al., 2021)

4.5 Simulation Result 4: False Positive Rates are Inflated when IL-HTE ($\sigma^2_{\zeta_1}$) is Ignored



JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

4.6 Simulation Result 4: False Positive Rates are Inflated when IL-HTE ($\sigma^2_{\zeta_1}$) is Ignored

- As a result of underestimated SEs, false positive rates increase in the constant effect model when IL-HTE is high.
- **Hypothesis tests may therefore be invalid** when IL-HTE is present but ignored in the model.

5 Empirical Application: The Model of Reading Engagement (MORE) Study

JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

5.1 Design

- Cluster randomized trial of 110 schools and 7797 students
- All students received a 12-day science lesson sequence
- Control students received a **double dose of science vocabulary lessons**
- Treatment students received **social studies extension lessons**

5.2 Research Question

- Could the treatment students leverage the schema for *system* and apply (transfer) it to new contexts?

5.2.1 Outcome Measure

- Reading comprehension assessment administered digitally
- 3 passages (Near, Mid, and Far transfer) with 10 items each

5.3 Research Hypotheses

- We hypothesized that there would be no treatment effect on the near transfer passage because it contained only the **science content**, whereas the mid and far transfer passages contained **social studies content** and would be more susceptible to the intervention.
- The mid transfer passage in particular might benefit from the treatment because the social studies extension lesson on the **Moon Mission** was very similar to the passage topic of the **Mars Mission.**

5.4 Assessment Excerpt from the "Mid Transfer" Passage

One of Earth's closest neighbors is the planet Mars. Scientists are curious about Mars. Is it like Earth? Is there life there? Could people live there? People want to learn about Mars. No one has made the long voyage there. Not even Neil Armstrong has explored Mars.

People need to overcome many obstacles to explore Mars.

First, astronauts need to be adventurous. It takes nearly 9 months to get to Mars. Astronauts need to invent new ways of eating in space. They need to experiment with ways to grow fruits and vegetables in space during the journey to Mars.

5.5 Example Question (Mid Transfer)

According to the passage, why do people's bones and muscles weaken?

1. Scientists are not sure why this happens
2. The journey from Earth to Mars takes one year
3. There is no oxygen in space to breathe
4. The pull of gravity on your body gets weaker

5.6 Modeling Strategy

Analogous models to the simulation, added random effect for school and standardized pretest reading score for precision. Strategy:

1. Constant Effect (Random Intercepts)
2. Randomly Varying Effect (IL-HTE)
3. Main effects and interactions for subtest passage (Near, Mid, Far)
4. Remove residual IL-HTE to determine if interactions explain all IL-HTE

5.7 MORE Model 3

$$\text{logit}(P(y_{ijk} = 1)) = \beta_0 + \beta_1 \text{treat}_k + \beta_2 \text{pretest}_{jk} + \beta_3 \text{mid}_i + \beta_4 \text{far}_i +$$
$$\beta_5 \text{treat}_k \times \text{mid}_i + \beta_6 \text{treat}_k \times \text{far}_i +$$
$$\theta_{jk} + \nu_k + \zeta_{0j} + \zeta_{1j} \text{treat}_k$$

5.8 Model Results

MORE EIRM Results

ATE is not significant in the constant effect model. Without considering IL-HTE, the analysis might end here!

	Mod. 1		Mod. 2		Mod. 3		Mod. 4	
Predictors	*Coef.*	*SE*	*Coef.*	*SE*	*Coef.*	*SE*	*Coef.*	*SE*
Intercept	0.16	0.13	0.16	0.13	0.34	0.19	0.34	0.19
1 = Treatment	0.06	0.09	0.06	0.09	0.03	0.09	0.03	0.09
Pretest (Std.)	0.91 ***	0.01	0.91 ***	0.01	0.91 ***	0.01	0.91 ***	0.01
1 = Mid Passage					-0.38	0.26	-0.38	0.26
1 = Far Passage					-0.16	0.25	-0.16	0.26
1 = Treatment x Mid					0.10 ***	0.03	0.10 ***	0.02
1 = Treatment x Far					0.00	0.03	0.00	0.02
Random Effects								
σ^2	3.2899		3.2899		3.2899		3.2899	
τ_{00}	0.4527 $_{s\_id}$		0.4528 $_{s\_id}$		0.4528 $_{s\_id}$		0.4528 $_{s\_id}$	
	0.1895 $_{sch\_id}$		0.1896 $_{sch\_id}$		0.1896 $_{sch\_id}$		0.1898 $_{sch\_id}$	
	0.3435 $_{item\_lab}$		0.3428 $_{item\_lab}$		0.3179 $_{item\_lab}$		0.3246 $_{item\_lab}$	
τ_{11}			0.0023 $_{item\_lab.s\_itt\_2122}$		0.0003 $_{item\_lab.s\_itt\_2122}$			
ρ_{01}			0.0055 $_{item\_lab}$		0.6676 $_{item\_lab}$			

The treatment effect on the **mid transfer items** is significantly stronger than the near or far items (B = 0.10 logits, p < 0.001), providing support for the transfer hypothesis

The interaction terms explain nearly all of the IL-HTE (87%)

*The EIRM provides a more **fine-grained** approach to HTE analysis by looking **within** the outcome measure*

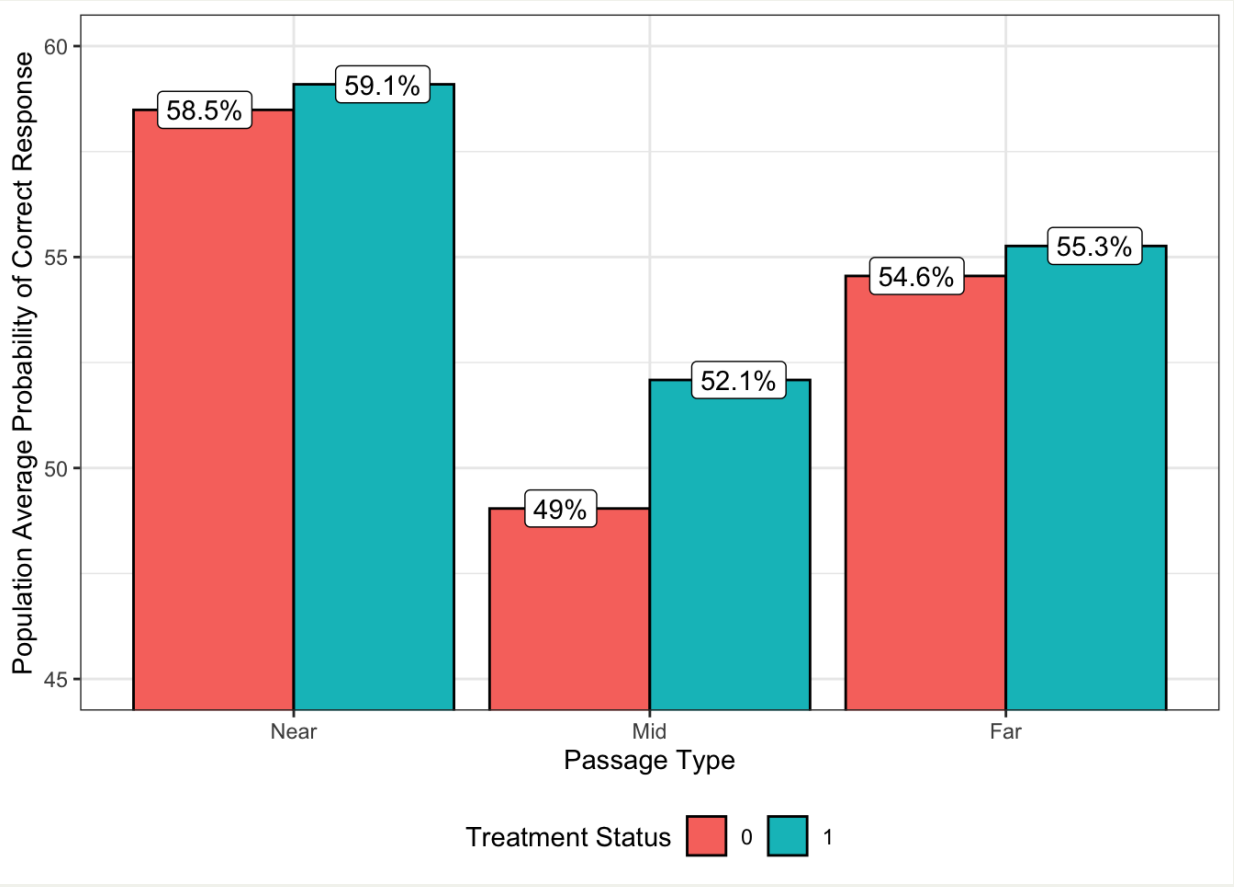
There is significant IL-HTE in the data - the item-level TE has an SD of about 0.05 (p<0.05)

JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

5.9 Visualizing the IL-HTE in Model 2

5.10 Visualizing the Passage-Level Treatment Effects from Model 4



JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

6 Summary and Conclusions

6.1 Summary and Conclusions

- HTE analysis methods have typically emphasized the **person level** (e.g., interactions)
- With the EIRM, we can examine HTE **within an outcome** by including a random slope for treatment at the item level
- When IL-HTE is present but ignored in the model, SEs can be underestimated and false positive rates can increase

6.2 Summary and Conclusions

- Failing to consider IL-HTE with the MORE data would have resulted in an incomplete conclusion about the efficacy of MORE on reading comprehension
- Modeling IL-HTE can provide more **fine-grained insight** for researchers on the efficacy of interventions, allowing us to better understand on which items or **types of items** treatment effects emerge, which is critical for policy relevance
- Recent work by Ahmed et al. (2023) suggests that IL-HTE is widespread in RCTs: "Based on our analysis of 7,244,566 item responses … taken from 15 RCTs … we find clear evidence for variation in gains across items."

7 Questions?

Thank you!

Acknowledgements: Jimmy Kim, Luke Miratrix, Doug Mosher, Jackie Relyea, Andrew Ho, HGSE READS Lab, HGSE Measurement Lab

joshua_gilbert@g.harvard.edu

josh.b.gilbert@gmail.com

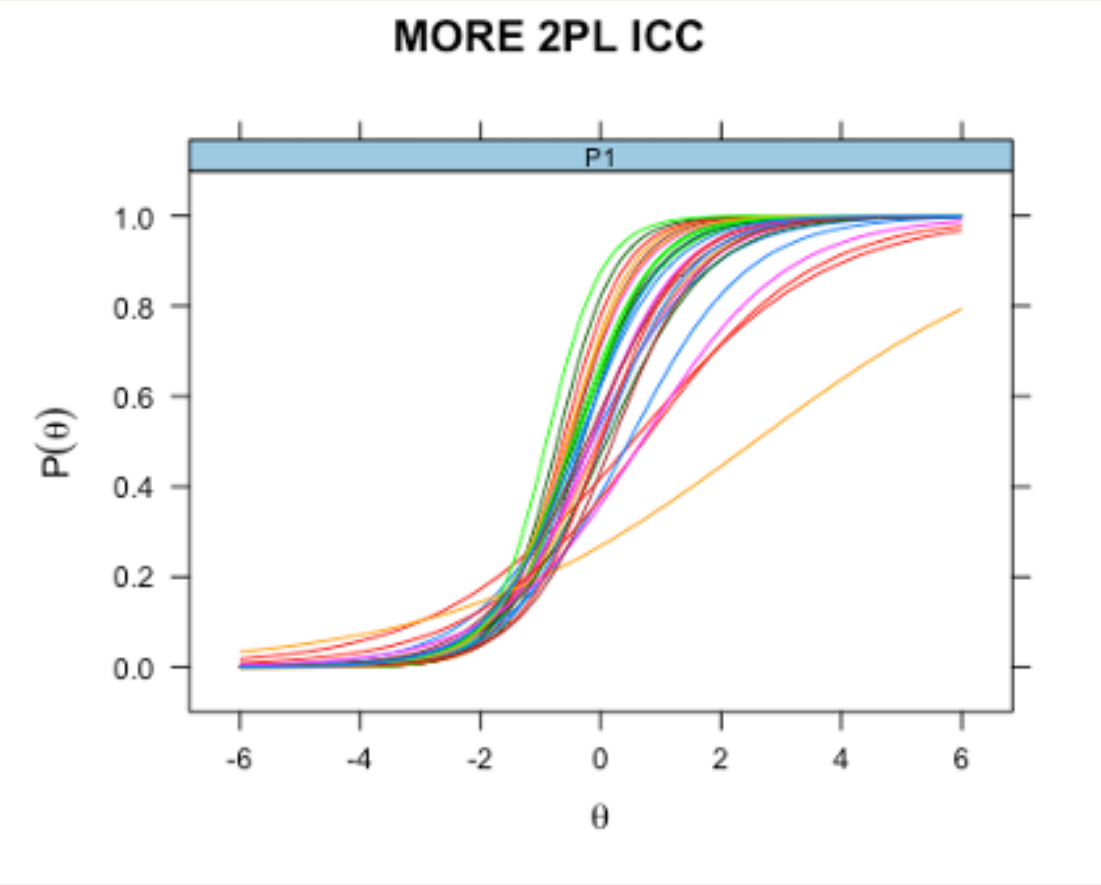
8 Appendices

8.1 The MORE Assessment is Unidimensional

JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

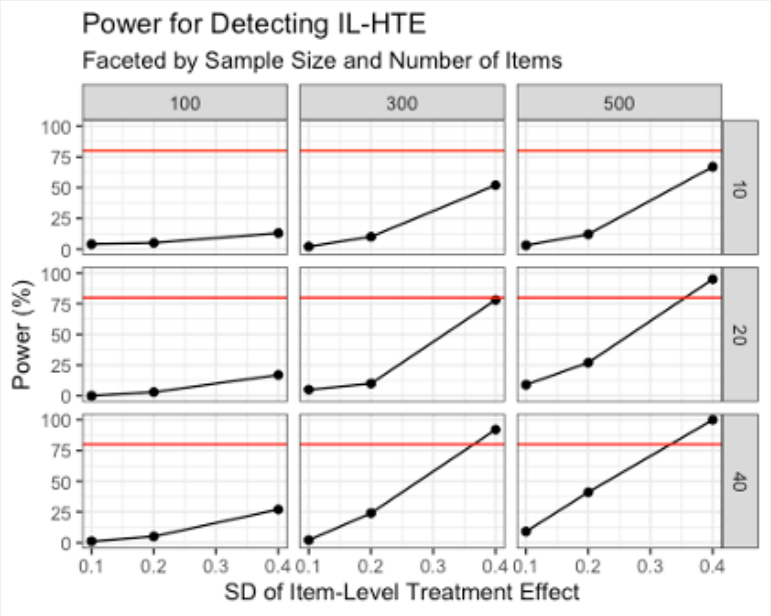
8.2 IRT ICC Curves for the MORE Assessment



JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB

8.3 Detecting IL-HTE ($\sigma^2_{\zeta_1}$) Requires Large Samples



JEBS IL-HTE Presentation (Gilbert, 2023)

READS LAB