ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-881

Disentangling Person-Dependent and Item-Dependent Causal Effects: Applications of Item Response Theory to the Estimation of Treatment Effect Heterogeneity

Joshua B. Gilbert
Harvard University

Luke W. Miratrix
Harvard University

Mridul Joshi
Stanford University

Benjamin W. Domingue
Stanford University

Analyzing heterogeneous treatment effects (HTE) plays a crucial role in understanding the impacts of educational interventions. A standard practice for HTE analysis is to examine interactions between treatment status and pre-intervention participant characteristics, such as pretest scores, to identify how different groups respond to treatment. This study demonstrates that identical patterns of HTE on test score outcomes can emerge either from variation in treatment effects due to a pre-intervention participant characteristic or from correlations between treatment effects and item easiness parameters. We demonstrate analytically and through simulation that these two scenarios cannot be distinguished if analysis is based on summary scores alone. We then describe a novel approach that identifies the relevant data-generating process by leveraging item-level data. We apply our approach to a randomized trial of a reading intervention in second grade, and show that any apparent HTE by pretest ability is driven by the correlation between treatment effect size and item easiness. Our results highlight the potential of employing measurement principles in causal analysis, beyond their common use in test construction.

VERSION: February 2024

Suggested citation: Gilbert, Joshua B., Luke W. Miratrix, Mridul Joshi, and Benjamin W. Domingue. (2024). Disentangling Person-Dependent and Item-Dependent Causal Effects: Applications of Item Response Theory to the Estimation of Treatment Effect Heterogeneity. (EdWorkingPaper: 23-881). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/6b7w-vp07

Disentangling Person-Dependent and Item-Dependent Causal Effects: Applications of Item Response Theory to the Estimation of Treatment Effect Heterogeneity

Joshua B. Gilbert*[1], Luke W. Miratrix[1], Mridul Joshi[2], and Benjamin W. Domingue[2]

[1]Harvard University Graduate School of Education
[2]Stanford University Graduate School of Education

Abstract

Analyzing heterogeneous treatment effects (HTE) plays a crucial role in understanding the impacts of educational interventions. A standard practice for HTE analysis is to examine interactions between treatment status and pre-intervention participant characteristics, such as pretest scores, to identify how different groups respond to treatment. This study demonstrates that identical patterns of HTE on test score outcomes can emerge either from variation in treatment effects due to a pre-intervention participant characteristic or from correlations between treatment effects and item easiness parameters. We demonstrate analytically and through simulation that these two scenarios cannot be distinguished if analysis is based on summary scores alone. We then describe a novel approach that identifies the relevant data-generating process by leveraging item-level data. We apply our approach to a randomized trial of a reading intervention in second grade, and show that any apparent HTE by pretest ability is driven by the correlation between treatment effect size and item easiness. Our results highlight the potential of employing measurement principles in causal analysis, beyond their common use in test construction.

Keywords: causal inference, heterogeneous treatment effects, item response theory, psychometrics, educational measurement

Forthcoming in the *Journal of Educational and Behavioral Statistics*

*Corresponding author: joshua_gilbert@g.harvard.edu

1 Introduction

Heterogeneous Treatment Effects (HTE) are a topic of growing importance and interest as understanding HTE is crucial for determining the policy relevance of interventions. HTE analysis allows researchers to identify the target population for which interventions are most effective. Further, identifying subgroups for whom there might be much larger or smaller treatment effects can lead to efficiency gains if policies are targeted to groups that are most responsive to interventions (Abenavoli, 2019; Blundell et al., 2005; Brand et al., 2014; Bryan et al., 2021; Schochet et al., 2014; Torche et al., 2024).

One challenge with estimating HTE is the risk that researchers may engage in questionable research practices such as selectively reporting subgroups with large effects or p-hacking, thus providing misleading conclusions about the generalizability and replicability of HTE analyses (Schuetze & von Hippel, 2023). Many methods have been proposed to address these problems, such as pre-registration (Olken, 2015) and machine learning approaches (Chernozhukov et al., 2022; Wager & Athey, 2018; Wallace et al., 2023; Yeager et al., 2019). These approaches address spurious HTE as a problem of inference. However, even when a researcher takes a principled approach to estimating HTE or employs novel machine learning methods, we show that they may still detect spurious HTE if the relevant outcome is a psychological construct or other form of latent outcome measured using psychometric techniques. This spurious HTE can occur when the items used to construct outcome measures, such as a test of reading comprehension based on a set of test items, themselves exhibit HTE. The present study thus complements the broader HTE literature by reconsidering spurious HTE as a problem of identification instead of a problem of inference. In particular, we show that two qualitatively different data-generating processes (DGPs) can yield identical patterns of HTE when psychometric outcomes such as educational test scores are used to estimate treatment effects.

To illustrate this problem, consider a common approach to estimating treatment effects. Education researchers often collect item-level assessment data, generate test scores for each individual from the item responses, and use these scores as outcomes in a standard regression framework to estimate treatment effects. The regression model can then be extended to include interactions between person characteristics and treatment status to probe potential sources of HTE. However, recent scholarship has proposed models where the treatment might differentially impact the assessment items (Ahmed et al., 2023; Gilbert et al., 2023b; Sales et al., 2021). Thus, by not modeling individual item-level responses directly, researchers are potentially ignoring the HTE in the items that constitute the outcome measure.

We offer two examples that could give rise to treatment effects that vary at the item level. First, in education, a treatment effect may reflect "teaching to the test" or "score inflation" rather than an improvement in the latent ability giving rise to the item responses (Koretz, 2005). For example, instruction in test-taking strategies such as "process of elimination" could have the effect of making multiple choice items uniformly easier without improving latent student ability. Second, in psychology, spousal loss induces a systematic change in the reporting of depressive symptoms such that people with otherwise similar levels of depression are much more likely to report the specific symptoms of loneliness and sadness, as opposed to other symptoms such as a loss of motivation (see Figure S7 in B. W. Domingue et al., 2021). Interpretation of this change for the purposes of diagnosis has been a long-standing challenge (e.g., Olivera-Aguilar and Rikoon, 2024; Zisook et al., 2007). These illustrations are not meant to be exhaustive but rather suggestive of the possibilities that we have in mind when we discuss item-dependent HTE.

In this study, we show that correlations between item-specific treatment effects and item easiness parameters can generate observed HTE patterns indistinguishable from those generated by HTE arising from treatment interacting with person characteristics. We demonstrate this resulting confounding both analytically and through Monte Carlo simulation. To resolve this issue, we propose a novel approach that leverages item-level data to simultaneously

estimate treatment by person characteristic interactions as well as the correlation between item-specific treatment effects and item easiness parameters. We show that our approach identifies the relevant DGP using Monte Carlo simulation. Finally, we apply our approach to a randomized evaluation of a second grade content literacy intervention and show that the observed HTE by pretest scores is driven by the correlation between item-specific treatment effects and the item easiness parameters.

Our study contributes to the burgeoning literature on the estimation of HTE (Athey & Imbens, 2016; Bryan et al., 2021; Chernozhukov et al., 2022; Künzel et al., 2019; Schuetze & von Hippel, 2023; Wager & Athey, 2018; Wendling et al., 2018). In particular, our study extends a literature on the identification and estimation of a latent heterogeneity—that is, the variation in treatment effects that is not driven by baseline variables (Jeon et al., 2021; Lyu et al., 2023; Pearl, 2022; VanderWeele & Batty, 2023; Winship & Morgan, 1999; Xie et al., 2012). We build on recent work describing how item-level assessment data, typically only used to construct outcome measures, can provide additional insights into the nature of treatment effects (Ahmed et al., 2023; Gilbert et al., 2023b; Sales et al., 2021), and more generally, how estimates of treatment impact may be sensitive to measurement properties such as the alignment between interventions and outcomes (Francis et al., 2022), the comparability of effect sizes across studies (Wolf & Harbatkin, 2023), effect moderation and mediation (Montoya & Jeon, 2020; VanderWeele & Vansteelandt, 2022), and the consequences of scoring decisions and outcome metric properties for inference (B. W. Domingue et al., 2020, 2022; Gilbert, 2023a; McNeish & Wolf, 2020; Skrondal & Laake, 2001; Soland et al., 2022; Widaman & Revelle, 2023). Our study leverages a measurement model (Briggs, 2008; De Boeck, 2004) within a potential outcomes framework (Holland, 1986; Rubin, 1974) to identify and estimate HTE, effectively integrating the three foundational elements of empirical research—measurement, identification, and inference—into a cohesive framework.

The study is structured as follows. Section 2 sets up a potential outcomes framework for the causal inference model and reviews person-dependent HTE and item-dependent HTE

DGPs. Section 3 examines how the two DGPs can yield identical observed patterns of heterogeneity, and our proposed solution to this problem. Section 4 presents a Monte Carlo simulation study, demonstrating the identification challenge and its resolution. Section 5 presents an empirical application to a randomized evaluation of a content literacy intervention. Section 6 discusses the implications of applying measurement principles in impact evaluation in education and other fields.

2 A Model of Treatment Effects

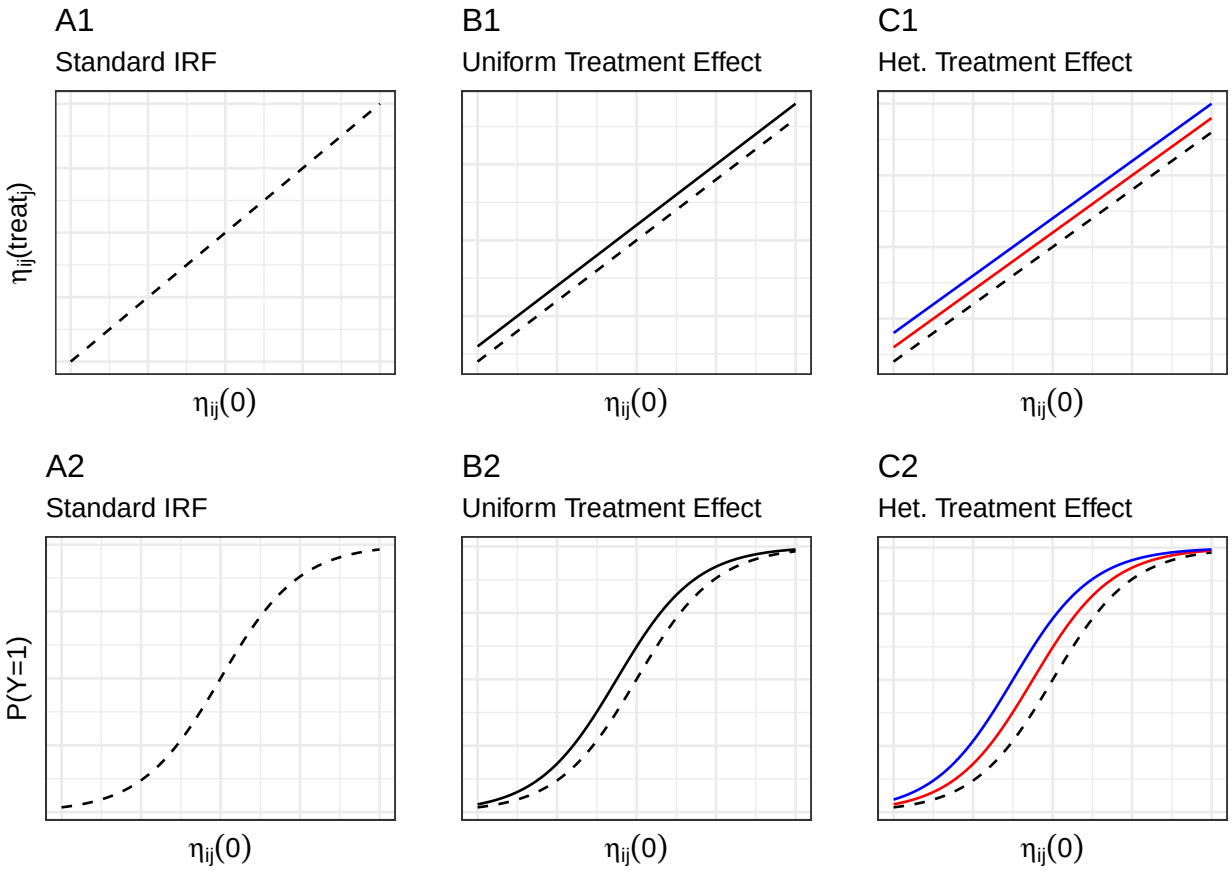
We begin with a model of treatment effects for item response data. We consider the following DGP for dichotomous item responses on a test used to evaluate the efficacy of a given treatment. The probability of a correct response to item i by person j is

$$\Pr(y_{ij} = 1) = f(\eta_{ij}) = f(\theta_j^1 + b_{ij}), \tag{1}$$

where f is a monotonically increasing function bounded within $(0, 1)$. Common choices for f include the inverse logit (logit^{-1}) and inverse probit (Φ^{-1}) functions, following a standard Rasch item response model (Baker, 2001; Embretson & Reise, 2013), or more flexible non-parametric approaches (Sijtsma, 2005). θ_j^1 denotes person j's post-intervention ability—we denote pre-intervention ability as θ_j^0—and b_{ij} is a parameter that indexes item easiness and that may vary across people. Our use of b as item easiness is the negative of an item's "difficulty" in the IRT literature. A basic version of Equation 1 where θ_j^1 and b_{ij} are constants and $f(x) = \text{logit}^{-1}(x)$ is shown in Panels A1 and A2 of Figure 1.

Both θ_j^1 and b_{ij} (and therefore η_{ij}) may be functions of treatment status and other relevant parameters such as pre-intervention ability (θ_j^0). Using the potential outcomes framework (Holland, 1986; Rubin, 1974), we define the causal effect of treatment τ_{ij} as the difference in

Figure 1: Item response functions (IRFs) and treatment effects.



Notes: The plots in the top row present the linear predictor and the plots in the bottom row present the probability of correct response. A1 and A2 depict a standard IRF that describes the probability of a correct response as a function of the sum $\theta_j^1 + b_i$. In all plots, the dashed black line shows the probability of a correct response under the control counterfactual, or $\eta_{ij}(0)$. In B1 and B2, the solid black line shows a uniform treatment effect on the standard IRF. In C1 and C2, solid lines depict the HTE. The IRF curves represent potential difference in functioning across items after exposure to treatment, with the blue IRF more impacted by treatment than the red IRF.

η_{ij} under the counterfactual treatment and control conditions:

$$\tau_{ij} = \eta_{ij}(1) - \eta_{ij}(0). \tag{2}$$

Suppose that an individual is randomly assigned to treatment or control conditions; we write $\text{treat}_j = 1$ if person j is assigned to treatment. Consider a uniform treatment effect of magnitude β_1 so that the intervention leads to an increase in ability for treated individuals,

$$\theta_j^1 = \theta_j^0 + \beta_1 \text{treat}_j \tag{3}$$

$$b_{ij} = b_i, \tag{4}$$

which implies $\eta_{ij}(1) = \theta_j^0 + \beta_1 + b_i, \eta_{ij}(0) = \theta_j^0 + b_i$, and $\tau_{ij} = \beta_1$. The implications of this kind of treatment effect are shown in Panel B of Figure 1; the vertical shift in panel B1 and the horizontal shift in the IRF in B2 would be β_1 in this example. Note that the shift uniformly affects all items and treated students, though the vertical distances between the curves differ because of the non-linear transformation to the probability space; the treatment effect is uniform in the latent linear predictor η_{ij} space (Breen et al., 2018; Colnet et al., 2023; Mood, 2010).

As an alternative, if we suppose that the treatment effect is entirely related to change in the functioning of the items,

$$\theta_j^1 = \theta_j^0 \tag{5}$$

$$b_{ij} = b_i + \beta_1 \text{treat}_j \tag{6}$$

we similarly produce $\tau_{ij} = \beta_1$. The fact that item response models are not identified under such scale translations in single-timepoint, single-population analyses (San Martín, 2016) is a key component of our argument.

We now introduce HTE into the model. At the person level, we can allow the treatment effect to depend on person characteristic X_j (e.g., pretest scores, gender, SES) by introducing an interaction between treat_j and X_j:

$$\theta_j^1 = \theta_j^0 + \beta_1 \text{treat}_j + \beta_2 X_j + \beta_3 \text{treat}_j \times X_j \tag{7}$$

$$b_{ij} = b_i. \tag{8}$$

At the item level, we can similarly allow the treatment effect size to depend on item characteristic X_i (e.g., item content, item type, position of item in the test):

$$\theta_j^1 = \theta_j^0 \tag{9}$$

$$b_{ij} = b_i + \beta_1 \text{treat}_j + \beta_2 X_i + \beta_3 \text{treat}_j \times X_i. \tag{10}$$

We consider HTE in Panel C of Figure 1. The red and blue lines represent different increases in η_{ij} or $\Pr(y_{ij} = 1)$ for different values of $\eta_{ij}(0)$. One possible interpretation is that the treatment effect depends on student characteristics. For example, if X_j represents baseline ability, a lower-ability student represented by the red line shows some increase in their probability of a correct response but this was less of an increase than the one observed by the higher ability student represented by the blue line. An alternative—and mathematically equivalent given that they lead to equivalent changes in η_{ij}—model would be to assume that items are differentially sensitive to treatment. The red IRF would then capture the post-treatment functioning of an item that showed a relatively smaller effect as compared to the blue IRF. We now formalize these ideas, which we describe as person-dependent HTE or item-dependent HTE respectively and discuss their relevance for empirical investigations of HTE.

Person-Dependent HTE

We first consider a DGP in which HTE arises due to variation in the treatment effect as a function of pre-treatment ability (θ_j^0):

$$\theta_j^1 = \beta_0 + \beta_1 \text{treat}_j + \beta_2 \theta_j^0 + \beta_3 \text{treat}_j \times \theta_j^0 \tag{11}$$

$$b_i \sim N(0, \sigma_0). \tag{12}$$

τ_{ij} is a function of both β_1 and β_3:

$$\begin{aligned}
\tau_{ij} &= \eta_{ij}(1) - \eta_{ij}(0) && (13)\\
&= \left(\theta_j^1(1) + b_i\right) - \left(\theta_j^1(0) + b_i\right) && (14)\\
&= \left(\beta_0 + \beta_1 \text{treat}_j + (\beta_2 + \beta_3)\theta_j^0 + b_i\right) - \left(\beta_0 + (\beta_2)\theta_j^0 + b_i\right) && (15)\\
&= \beta_1 + \beta_3 \theta_j^0. && (16)
\end{aligned}$$

Here, β_3 is the person-dependent HTE parameter. Suppose β_3 is 0. In this case, the treatment effect is constant and independent of baseline ability. Further, the average treatment effect (ATE = $\overline{\tau_{ij}}$) is simply β_1. In contrast, if $\beta_3 < 0$ then there are larger treatment effects when $\theta_j^0 < 0$ (and reversed if $\beta_3 > 0$). We emphasize the fact that the heterogeneity in this model is person-dependent; specifically it is due to pre-intervention variation in ability. For identification purposes, we assume here that b_i are normally distributed with mean 0 and some standard deviation σ_0.

Item-Dependent HTE

We now contrast the person-dependent HTE model with an alternative DGP:

$$\theta_j^1 = \gamma_0 + \gamma_1 \text{treat}_j + \gamma_2 \theta_j^0 \tag{17}$$

$$b_{ij} = b_i + \zeta_i \text{treat}_j \tag{18}$$

$$\begin{bmatrix} b_i \\ \zeta_i \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_0 & \rho \\ \rho & \sigma_1 \end{bmatrix}\right). \tag{19}$$

Again consider the treatment effect τ_{ij}. In contrast to the person-dependent DGP, here, τ_{ij} is constant as a function of j and, in particular, invariant as a function of θ_j^0. However, items are differentially affected by treatment, represented by ζ_i:

$$\tau_{ij} = \eta_{ij}(1) - \eta_{ij}(0) \tag{20}$$

$$= (\gamma_0 + \gamma_1 + \gamma_2 \theta_j^0 + b_i + \zeta_i) - (\gamma_0 + \gamma_2 \theta_j^0 + b_i) \tag{21}$$

$$= \gamma_1 + \zeta_i. \tag{22}$$

The ζ_i parameter has been previously described as "item-level" HTE (IL-HTE), "treatment by item interactions", or "item-specific effects" (Ahmed et al., 2023; Gilbert et al., 2023b; Sales et al., 2021). For example, if $\zeta_i > 0$, then responses to item i are more likely to be correct if $\text{treat}_j = 1$, above and beyond the ATE γ_1 (assumed here to be impacting θ_j^1 directly), independent of θ_j^1. The ζ_i offsets are equivalent to uniform differential item functioning (DIF) effects caused by the treatment (Camilli, 2006; Gilbert et al., 2023b; Montoya & Jeon, 2020; Santelices & Wilson, 2010). In this model, we assume that b_i and ζ_i come from a multivariate normal distribution with mean 0 and standard deviations σ_0 and σ_1, and correlation ρ. ρ captures potential associations between item easiness and item-specific treatment effect size. For example, if $\rho > 0$, then easier items are more likely to be positively affected by treatment. The parameter ρ will play a key role in appropriate identification of HTE.

A Directed Acyclic Graph (DAG) representation of the two DGPs is presented in Figure 2. Concretely, the person-dependent HTE model may be theoretically justifiable when an intervention targets students by ability level. For example, an intervention that provides more resources and supports to lower ability students would presumably have a larger effect on low ability students and this effect would be constant across all items (i.e., $\beta_3 < 0, \sigma_1 = 0$). In contrast, the item-dependent HTE model may be theoretically justifiable when an intervention targets skills by complexity level. For example, an intervention that provides foundational skills training to all students would potentially help students at all levels of θ_j^0 equally but show the largest impact on the easiest test items (i.e., $\rho > 0, \beta_3 = 0$). A treatment that leads to greater gains for low-scoring students will have different policy implications than a treatment that leads to greater learning of easier content for all students.

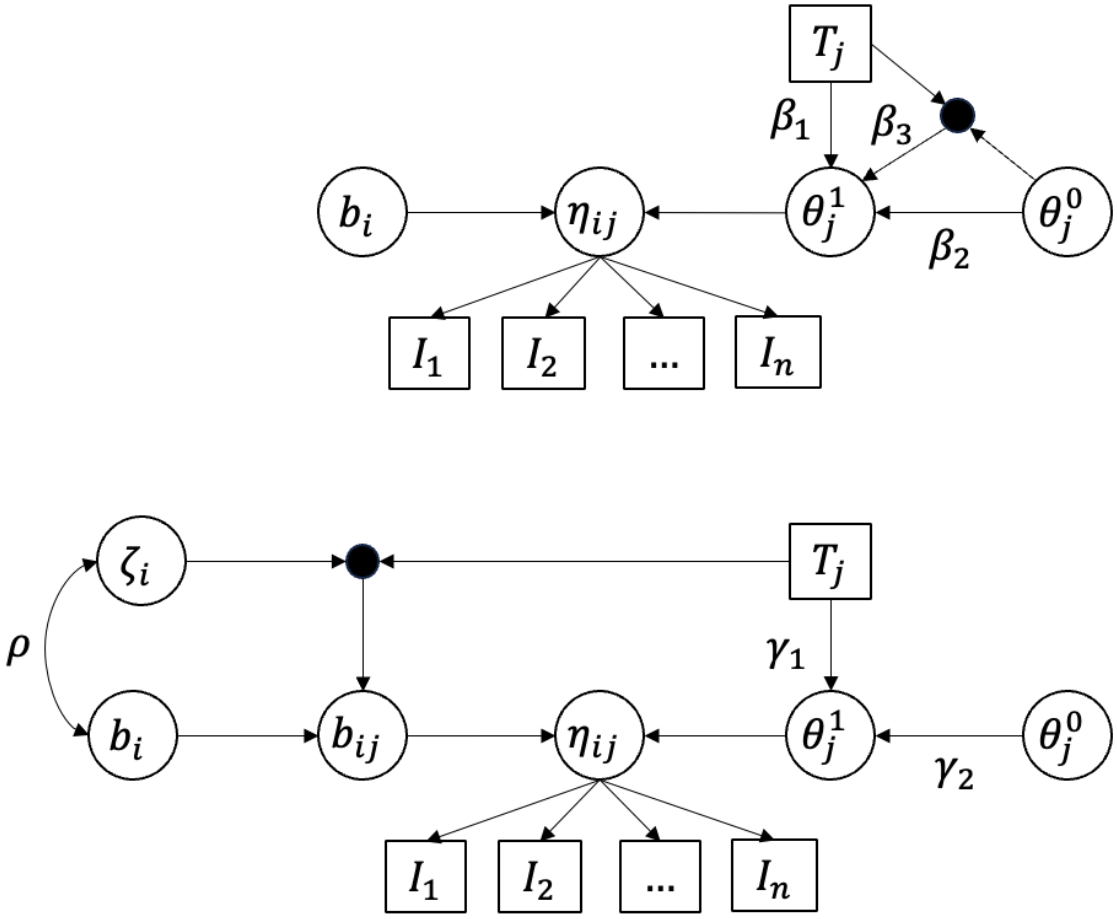
3 Appropriate Identification of HTE

The Problem of Identification

We now describe a scenario under which conventional analysis of HTE would leave the DGP unidentified. By this we mean that the person- or item-HTE processes produce data with equivalent patterns of treatment effects on an estimated post-treatment sum score S_j conditional on θ_j^0. Our argument hinges on the ρ parameter from Equation 19 which indexes the correlation between item easiness b_i and residual item-specific treatment effect size ζ_i.

We use sum scores for clarity of exposition and because—despite active debate among psychometric researchers concerning their properties (McNeish & Wolf, 2020; Soland et al., 2022; Widaman & Revelle, 2023)—they are commonly used as outcomes in empirical research (Flake et al., 2017). However, our argument is not dependent on the use of sum scores. When items are equally discriminating, the sum score is a sufficient statistic for an IRT-based score. In fact, sum scores will be an injective (i.e., one-to-one) function of the true abilities (Birnbaum, 1968; Borsboom, 2005), but monotonic transformations can induce HTE, or

Figure 2: Directed acyclic graphs for person-dependent HTE (top) and item-dependent HTE (bottom) DGPs



Notes: Squares indicate observed variables, hollow circles indicate latent variables, and solid circles represent cross product interaction terms. I_n are item responses and T_j is the treatment indicator. β_3 represents the treatment by baseline ability interaction. ρ represents the correlation between item easiness and item-specific treatment effect size. Path coefficients are fixed at 1 unless otherwise indicated.

remove it (Ding et al., 2019; B. W. Domingue et al., 2022). Further, our simulations will show that our argument applies equally to latent variable models that estimate the measurement and regression models simultaneously (Lockwood & McCaffrey, 2020).

Let $\mathbb{E}(S_j)$ be the sum of the IRFs. That is, $\mathbb{E}(S_j)$ is the test response function, or TRF, which can also be interpreted as the Classical Test Theory true score (Borsboom, 2005). We define the TRF for the person-dependent process as $\mathbb{E}_P(S_j)$ and the TRF for the item-dependent process as $\mathbb{E}_I(S_j)$:

$$\mathbb{E}_P(S_j) = \sum_{i=1}^{I} f(\eta_{ij}) = \sum_{i=1}^{I} f(\beta_0 + \beta_1 \text{treat}_j + \beta_2 \theta_j^0 + \beta_3 \text{treat}_j \times \theta_j^0 + b_i) \tag{23}$$

$$\mathbb{E}_I(S_j) = \sum_{i=1}^{I} f(\eta_{ij}) = \sum_{i=1}^{I} f(\gamma_0 + \gamma_1 \text{treat}_j + \gamma_2 \theta_j^0 + \text{treat}_j \times \zeta_i + b_i). \tag{24}$$

We define the causal estimand at the test score level as τ, where $S_j(\cdot)$ is the potential value of the sum score under the treatment or control counterfactuals:

$$\tau = \mathbb{E}\left[S_j(1) - S_j(0)\right] = \mathbb{E}(S_j(1)) - \mathbb{E}(S_j(0)). \tag{25}$$

We emphasize that τ is distinct from τ_{ij} described earlier in Section 2, in which treatment effects were defined in terms of η_{ij}. Here, τ is averaged across items and persons.

We first illustrate the identification problem with a toy example, setting $\beta_0 = 0$, $\beta_1 = 0$, $\beta_2 = 1$, $\gamma_0 = 0$, $\gamma_1 = 0$, and $\gamma_2 = 1$. In the person-dependent HTE model, we set $\beta_3 = -0.28$ and in the item-dependent HTE model we set $\rho = 1$ so that $\zeta_i = b_i$. Thus, the treatment is most effective for students with lower θ_j^0 under the person DGP (for all items), and the treatment is most effective for the easiest items under the item DGP (for all persons). Further assuming $f() = \text{logit}^{-1}$ (i.e., a Rasch or 1PL IRT model), the equations simplify to:

$$\mathbb{E}_P(S_j) = \sum_{i=1}^{I} \text{logit}^{-1}(\theta_j^0 - 0.28\text{treat}_j \times \theta_j^0 + b_i) \tag{26}$$

$$\mathbb{E}_I(S_j) = \sum_{i=1}^{I} \text{logit}^{-1}(\theta_j^0 + \text{treat}_j \times b_i + b_i). \tag{27}$$

In Table 1, we calculate $\mathbb{E}(S_j)$ for three items with $b_i = -1, 0, 1$ and three individuals with $\theta_j^0 = -1, 0, 1$ under both treatment and control counterfactuals. We see that $\mathbb{E}(S_j)$ are identical under both DGPs which implies that τ is identical. We also see that τ is a linear function of θ_j^0 under both DGPs, with the largest τ emerging for the lowest values of θ_j^0 even though there is no treatment by θ_j^0 interaction in the item-HTE DGP. As we will show, without item-level outcome data supplementing pre-intervention data, there is no way to distinguish between these two competing causes of HTE.

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

Notes: Toy example shows how both the person-dependent and item-dependent DGPs produce treatment effects on sum scores that depend on θ_j^0. f is a logit^{-1} function. $\mathbb{E}(S_j)$ is calculated for three items with $b_i = -1, 0, 1$ and three individuals with $\theta_j^0 = -1, 0, 1$ under both treatment and control conditions. We set $\beta_0 = 0$, $\beta_1 = 0$, $\beta_2 = 1$, $\gamma_0 = 0$, $\gamma_1 = 0$, and $\gamma_2 = 1$. In the person HTE model, we set $\beta_3 = -0.28$ and in the item HTE model we set $\rho = 1$ so that $\zeta_i = b_i$. We calculate τ by subtracting the expected scores under control condition from expected scores under treatment condition.

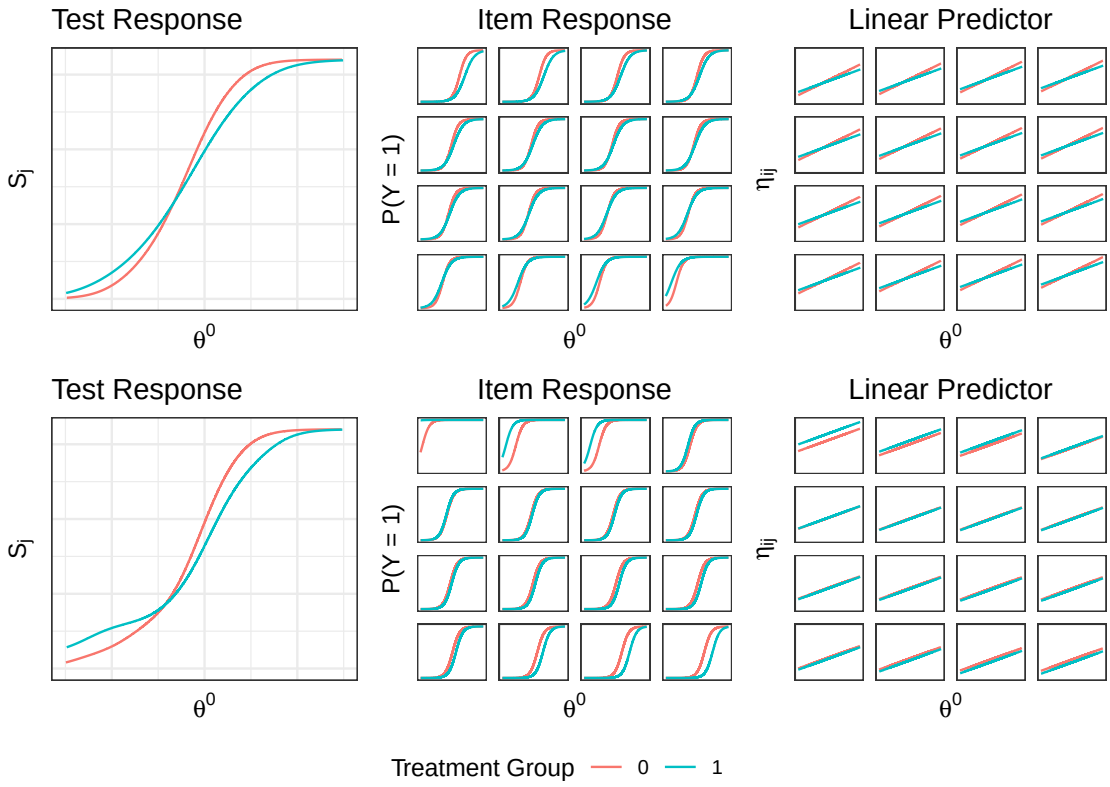
This phenomenon holds in more general settings. For example, Figure 3 shows a more realistic case with 16 items, where the top row depicts a person-dependent HTE scenario where $\beta_3 < 0$ and the bottom row depicts an item-dependent HTE scenario where $\rho = 1$. The first column shows $\mathbb{E}(S_j)$ as a function of θ_j^0 by summing the IRFs for all 16 items, and we can see that the pattern is essentially identical in both DGPs, similar to the result in

Table 1. That is, given just $\mathbb{E}(S_j)$, the two DGPs are empirically indistinguishable (Spoto & Stefanutti, 2023), which provides a serious interpretational problem given the common usage of psychometric outcomes in RCTs and the typical analysis focusing on test-level aggregates. Using the items, however, may help resolve this problem. The second and third columns show the individual IRFs on the probability scale and linear predictor η_{ij} scale, respectively. When examining the item-level data, we can clearly distinguish the DGPs. In the person-HTE model, the IRFs cross for each item, and summing the IRFs creates the overall crossing pattern in $\mathbb{E}(S_j)$. In the item-HTE model, the IRFs are parallel for each item, but the relative vertical distance between the groups varies across items, creating the identical crossing pattern in $\mathbb{E}(S_j)$. The stark distinction between IRFs for each DGP suggest that these two processes can potentially be distinguished when the appropriate model is applied to item-level data.

We can formalize the intuition provided by Table 1 and Figure 3 analytically. We begin with an arbitrary set of β for the person-HTE model and find a set of γ from the item-HTE model that will yield identical τ. While there is no simple formulation of $\mathbb{E}_P(S_j(\cdot))$ in general (Baker, 2001, pp. 69-70), we have assumed that items are equally discriminating and their easiness parameters are normally distributed, which enables a tractable solution. Under these assumptions, $\mathbb{E}_P(S_j(\cdot))$ can be approximated by a logistic curve scaled by the number of items I, as the sum of logistic curves can be approximated by a logistic curve when the slopes are equal and the curves have considerable overlap (Reed & Pearl, 1927, p. 733). In fact, our simulations show that the logistic approximation of $\mathbb{E}(S_j)$ is incredibly accurate, with an R^2 of greater than 99.9%.[1] We can therefore approximate the $\mathbb{E}_P(S_j(\cdot))$ follows:

[1]Summing logisitic curves is conceptually similar to constructing population average logistic curves using Generalized Estimating Equations from cluster-specific logistic curves derived from multilevel models. Thus, the slope of the TRF will be attenuated compared the slopes of the IRFs by a known quantity. When $\sigma_0 = 1$, this quantity is about 1.16. This attenuation does not affect our argument because we could simply set the item discrimination to 1.16 to achieve the desired result. See Austin and Merlo, 2017, p. 3263, Rabe-Hesketh and Skrondal, 2022, pp. 586-590, and Neuhaus et al., 1991, p. 28 for a discussion.

Figure 3: Different DGPs can produce the same pattern of sum score responses



Notes: This figure illustrates how different DGPs can produce the same pattern of sum score responses. The top row depicts person-dependent HTE and the bottom row depicts item-dependent HTE. We first generated the item-level data from the item-dependent HTE model where we set $\rho = 1$, implying that the treatment effect is largest on the easiest items. We then fit the misspecified person-dependent HTE model to the data generated by the item-dependent HTE model, and used the parameter estimates from this model to generate the data for the top row of the figure.

$$\mathbb{E}_P(S_j(\cdot)) = \sum_{i=1}^{I} \text{logit}^{-1}(\theta_j^1(\cdot) + b_i) \approx \text{logit}^{-1}(\theta_j^1(\cdot))I. \tag{28}$$

Applying the above equation to each potential outcome yields the following logistic functions, each of which can in turn be approximated by a linear equation because logistic curves are approximately linear between about 20% to 80% of the upper asymptote (Long, 1997; Von Hippel, 2015), where $\frac{I}{2}$ represents the midpoint of the scale and $\frac{I}{4}$ represents the linear approximation of the slope at the midpoint in a Taylor expansion:

$$\mathbb{E}_P(S_j(1)) = \sum_{i=1}^{I} \text{logit}^{-1}(\beta_0 + \beta_1 + \beta_2\theta_j^0 + \beta_3\theta_j^0 + b_i) \tag{29}$$

$$\approx \text{logit}^{-1}(\beta_0 + \beta_1 + \beta_2\theta_j^0 + \beta_3\theta_j^0)I \tag{30}$$

$$\approx \frac{I}{2} + \frac{I}{4}(\beta_0 + \beta_1 + \beta_2\theta_j^0 + \beta_3\theta_j^0) \tag{31}$$

$$\mathbb{E}_P(S_j(0)) = \sum_{i=1}^{I} \text{logit}^{-1}(\beta_0 + \beta_2\theta_j^0 + b_i) \tag{32}$$

$$\approx \text{logit}^{-1}(\beta_0 + \beta_2\theta_j^0)I \tag{33}$$

$$\approx \frac{I}{2} + \frac{I}{4}(\beta_0 + \beta_2\theta_j^0). \tag{34}$$

The treatment effect, τ, is the difference in these quantities:

$$\tau_P \approx \frac{I}{4}(\beta_1 + \beta_3\theta_j^0). \tag{35}$$

For the item-HTE model, we assume $\rho = 1$ so that $\zeta_i = b_i$, yielding the following approximations for $\mathbb{E}_I(S_j(\cdot))$. The division by 2 in $\mathbb{E}_I(S_j(1))$ comes from the presence of two b_i terms.

$$\mathbb{E}_I(S_j(1)) = \sum_{i=1}^{I} \text{logit}^{-1}(\gamma_0 + \gamma_1 + \gamma_2\theta_j^0 + b_i + b_i) \tag{36}$$

$$\approx \text{logit}^{-1}(\frac{\gamma_0 + \gamma_1 + \gamma_2\theta_j^0}{2})I \tag{37}$$

$$\approx \frac{I}{2} + \frac{I}{8}(\gamma_0 + \gamma_1 + \gamma_2\theta_j^0) \tag{38}$$

$$\mathbb{E}_I(S_j(0)) = \sum_{i=1}^{I} \text{logit}^{-1}(\gamma_0 + \gamma_2\theta_j^0 + b_i) \tag{39}$$

$$\approx \text{logit}^{-1}(\gamma_0 + \gamma_2\theta_j^0)I \tag{40}$$

$$\approx \frac{I}{2} + \frac{I}{4}(\gamma_0 + \gamma_2\theta_j^0). \tag{41}$$

We again compute τ:

$$\tau_I \approx \frac{I}{8}(\gamma_1 - \gamma_0 - \gamma_2\theta_j^0). \tag{42}$$

Clearly, τ is a linear function of θ_j^0 in both DGPs, analogous to the visualization shown in Figure 3.[2]

We now have two expressions for τ, one for each DGP, both of which can be approximated by a linear function of θ_j^0. Thus, there will be no way of differentiating between the two DGPs if we can identify values of these functions that lead to equivalent expressions for τ in the above equations. By setting $\tau_P = \tau_I$, we see that the DGPs provide equivalent τ when

[2]An intriguing implication of these results is that an identical pattern would emerge if the treatment changed the discrimination parameter of the items. For example, we get the division by 2 from the addition of the second b_i term in the equation for $\mathbb{E}_I(S_j(1))$, but the same result would manifest if the treatment reduced the discrimination of the items by a factor of 2. However, such a treatment is hard to imagine empirically, so we do not pursue the possibility further.

$$\gamma_1 - \gamma_0 = 2\beta_1 \tag{43}$$

$$\gamma_2 = -2\beta_3 \tag{44}$$

as the γ are fixed but unknown constants. Thus, models for $\mathbb{E}(S_j)$ alone cannot distinguish between the two DGPs.

The Solution

We can identify the relevant DGP with item-level data by including *both* the baseline ability by treatment interaction β_3 and the correlation between item easiness and item-specific treatment effect size ρ in a single model, such as an Explanatory Item Response Model (EIRM). The EIRM is a special case of the generalized linear mixed model (GLMM) that allows for the simultaneous estimation of an IRT measurement model and an explanatory regression model with person- or item-level predictors in a single procedure (Briggs, 2008; De Boeck, 2004; De Boeck et al., 2011; Gilbert, 2023b, 2024; Gilbert et al., 2023b; Petscher et al., 2020).

We propose the following flexible EIRM to disentangle the causes of HTE:

$$\text{logit}(y_{ij} = 1) = \eta_{ij} = \theta_j^1 + b_{ij} \tag{45}$$

$$\theta_j^1 = \beta_0 + \beta_1 \text{treat}_j + \beta_2 \theta_j^0 + \beta_3 \text{treat}_j \times \theta_j^0 \tag{46}$$

$$b_{ij} = b_i + \zeta_i \text{treat}_j \tag{47}$$

$$\begin{bmatrix} b_i \\ \zeta_i \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_0 & \rho \\ \rho & \sigma_1 \end{bmatrix}\right). \tag{48}$$

Because this model allows for both person and item-dependent HTE and models the item-level data directly, we can determine the extent to which person or item HTE is better fit to the

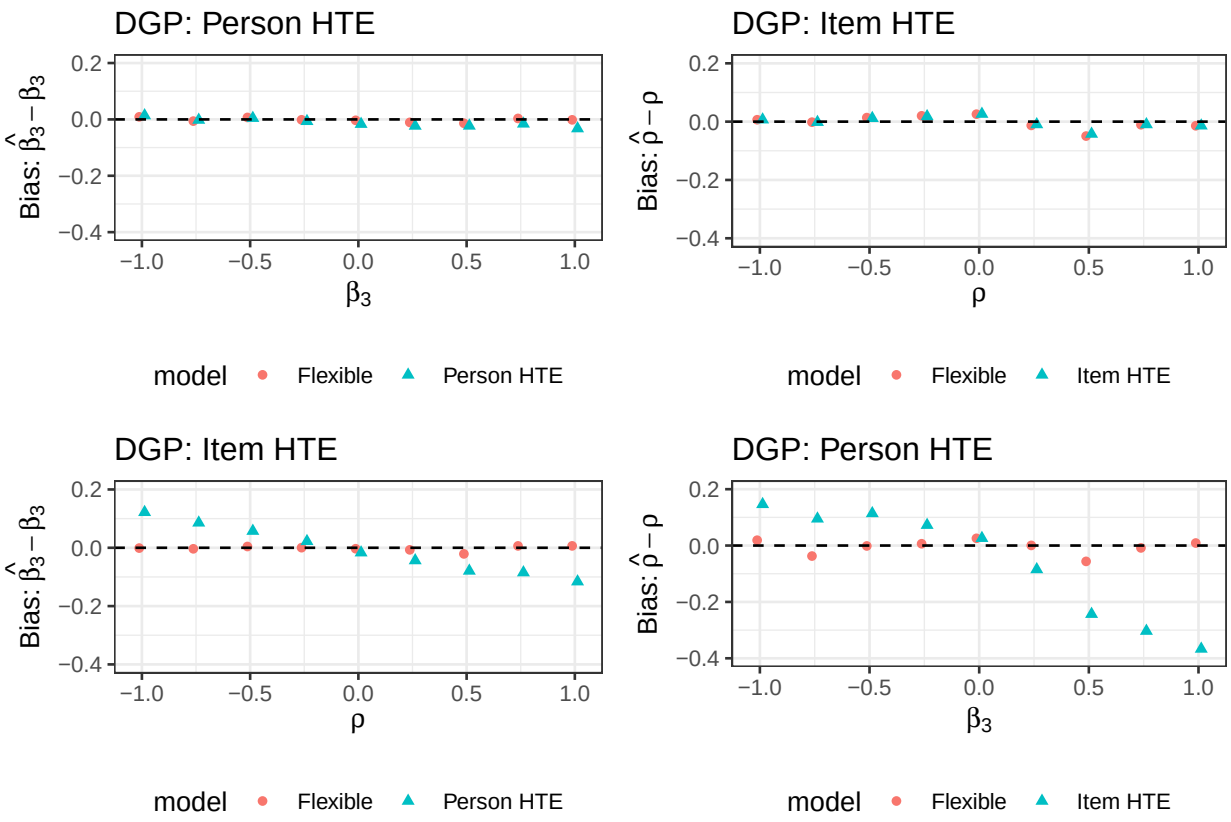
underlying item-level data and identify the appropriate causes of HTE. We turn now to a simulation study demonstrating this result.

4 Monte Carlo Simulation

A key implication of the empirical indistinguishability of the person-dependent and item-dependent HTE processes is that we will estimate a non-zero value of β_3 if ρ is the true data-generating parameter, and vice-versa. To demonstrate that person-dependent and item-dependent HTE can become confounded, we perform a Monte Carlo study across a range of β_3 and ρ values to demonstrate the effects of model misspecification across a range of more realistic testing conditions. To maintain focus on our two parameters of interest, we fix the following parameters across all trials: 500 subjects, 40 items with constant discrimination of 1, $\beta_1 = 0.20, \beta_2 = 1, \sigma_0 = 3, \sigma_1 = 0.5, \text{SD}(\theta_j^0) = 1$. We vary β_3 and ρ independently from -1 to 1 in increments of 0.25. We fit three EIRMs to each simulated data set, one with a treatment by baseline ability interaction β_3, the other with a ρ term, and the third with both β_3 and ρ, using the `glmer` function from the `lme4` package in R (i.e., a Rasch model) (De Boeck et al., 2011; Gilbert, 2023b), collect the output of the models for further analysis, and repeat the process 100 times for each combination of parameters.

The simulation results are shown in Figure 4. The plots in the first row show that each model appropriately recovers the target parameter with minimal bias. That is, whether the DGP includes treatment by baseline ability interaction β_3 or treatment by item easiness correlation ρ, $\hat{\beta}_3$ and $\hat{\rho}$ estimates from the appropriately specified model are unbiased, as are estimates from the flexible model that allows for both β_3 and ρ. The plots in the second row show the consequences of model misspecification. We observe that a non-zero data-generating value of ρ causes bias in $\hat{\beta}_3$ and vice versa. In contrast, when we fit the flexible model that allows for both β_3 and ρ, we can see that the bias is removed. Therefore, when we observe an apparent treatment by pretest interaction a test score outcome, we cannot determine whether

Figure 4: Simulation results showing bias in parameter estimates



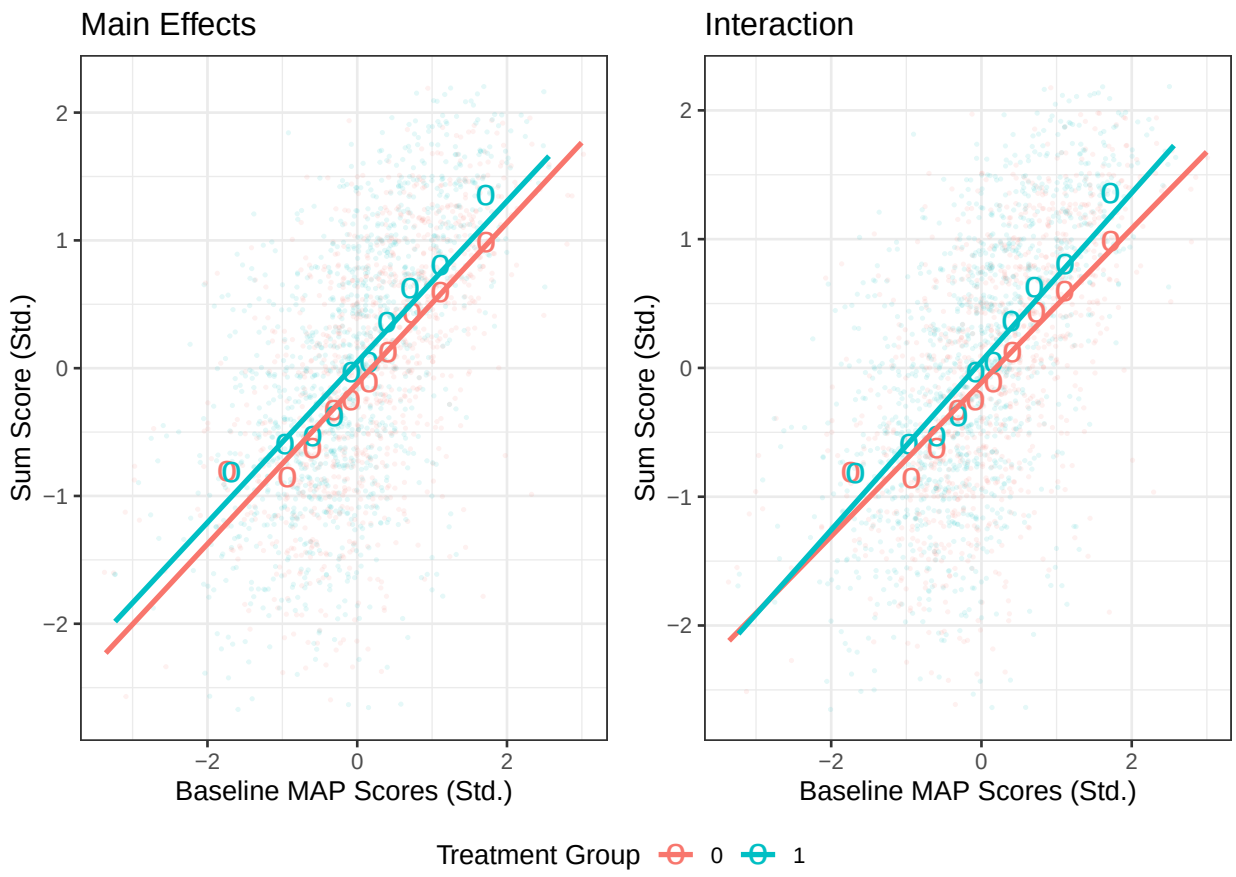
Notes: Figures depict bias in parameter estimates. The top row presents properly specified models and the bottom row presents misspecified models. The following parameters are fixed across all trials: 500 subjects, 40 items with constant discrimination of 1, $\beta_1 = 0.20, \beta_2 = 1, \sigma_0 = 3, \sigma_1 = 0.5$. β_3 and ρ are varied independently from -1 to 1 in increments of 0.25.

this is due to true person heterogeneity or a correlation between item treatment effect size and item easiness, unless we model the item-level data directly.

5 Empirical Application

For our empirical application, we analyze a public use data set from Kim et al., 2022, which examined the impact of the Model of Reading Engagement (MORE) literacy intervention on the reading comprehension test scores of 2174 grade 2 students. The researcher-designed reading comprehension test contained 20 dichotomous items. Figure 5 shows the standardized

Figure 5: Scatter plot of standardized sum-score outcome on pre-intervention scores

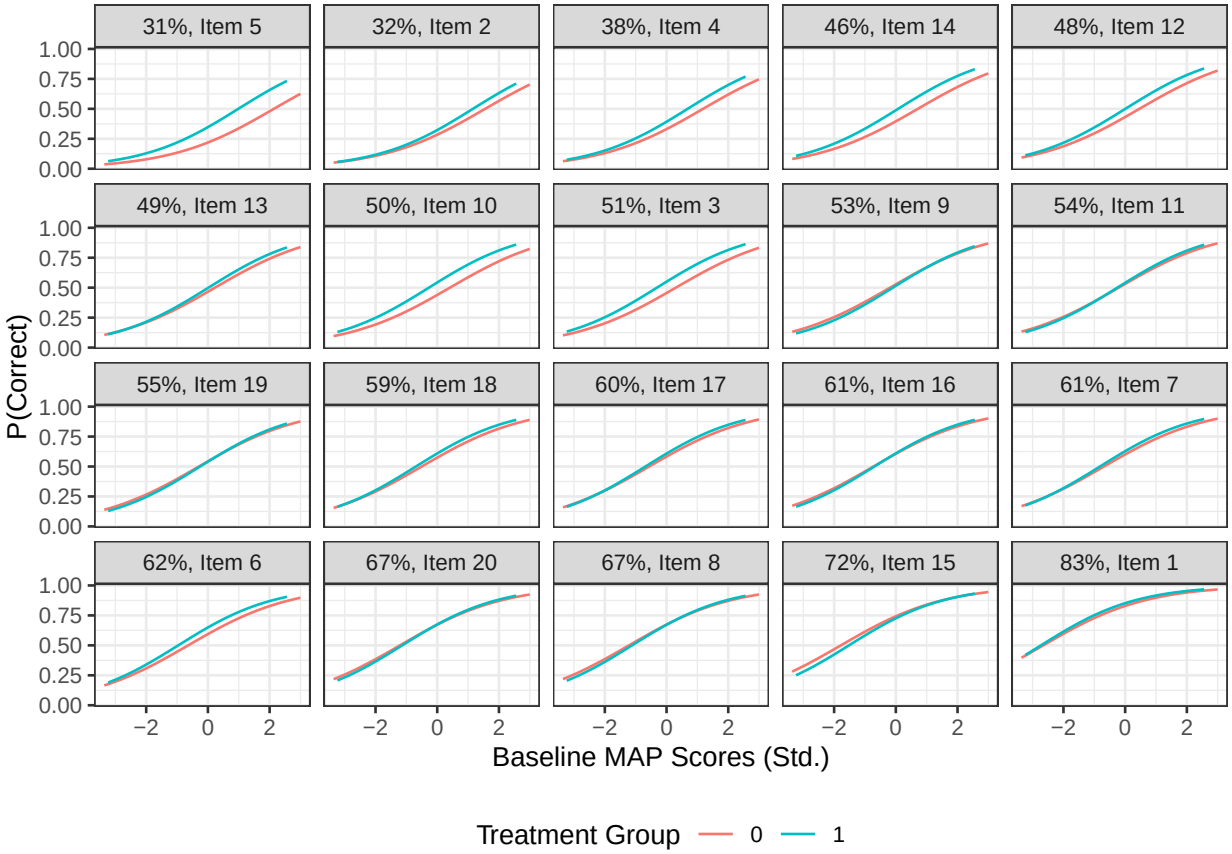


Notes: The figure presents a scatter plot of standardized reading comprehension sum scores on pre-intervention Measure of Academic Progress (MAP) reading scores by treatment status. The left panel depicts the direct effects of the treatment and the right panel also includes the interaction effect. Hollow circles indicate the conditional means of the outcome at each decile of the baseline score.

outcome sum score on the pretest score, where the right panel appears to show that the treatment effect is larger for students with higher pre-intervention Measure of Academic Progress (MAP) scores.

However, it would be premature to assert that the treatment is more effective for students with higher pre-intervention achievement (i.e., $\beta_3 > 0$) without conducting an item-level analysis; the data are equally consistent with a treatment that produces the largest impacts on the most difficult items (i.e., $\rho < 0$). As such, we replicate the middle panel of Figure 3 using our empirical item-level data in Figure 6. Visually, it appears that the treatment effects

Figure 6: Fitted Logistic Curves by Treatment Group for each Item



Notes: The figure presents fitted logistic curves of the probability of a correct response on pre-intervention Measure of Academic Progress (MAP) reading scores by treatment status. The items are arranged by difficulty, with percent correct values listed in the headings. The logistic curves are derived from a fixed effects model of the correct response as a function of treatment, pretest, and item indicator, with two-way interactions between treatment and pretest and treatment and the item indicators.

are largest on the most difficult items, rather than the treatment showing the strongest effect on high achieving students across all items.

To formally test this hypothesis and determine the source of HTE, we fit four EIRMs to the item-level data. Our baseline specification is a constant treatment effect model as described in Equation 3. We then fit a person-dependent HTE model that incorporates an interaction between treatment and pretest scores, β_3, as described in Equation 11 and an item-dependent HTE model that includes a correlation between item easiness and treatment

effect size, ρ, as described in Equation 18. Our final specification fits the flexible model, integrating both β_3 and ρ, as specified in Equation 45. The results are displayed in Table 2.

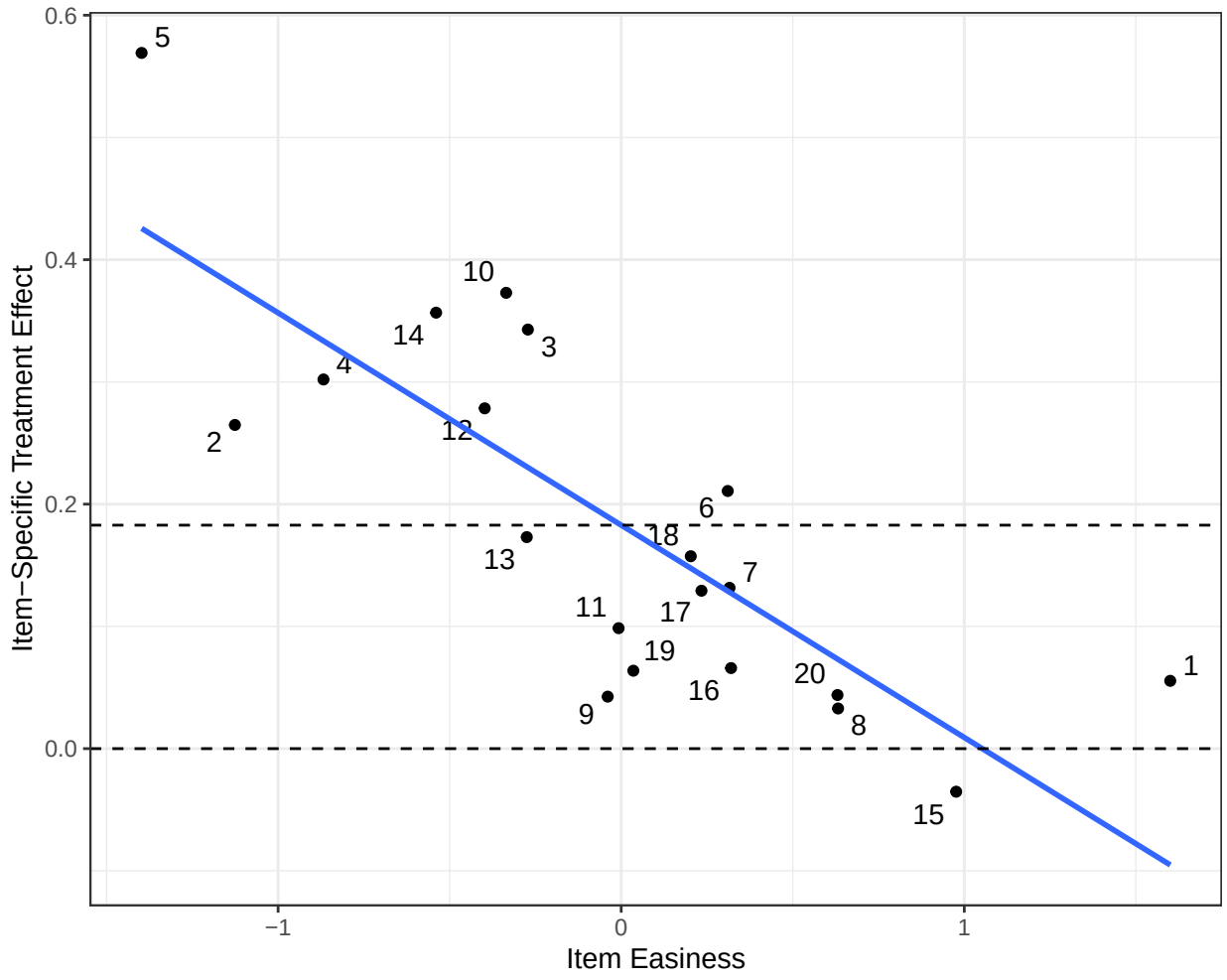
Table 2: Empirical Application of Explanatory Item Response Models

	(1)	(2)	(3)	(4)
Constant	0.12 (0.15)	0.12 (0.15)	0.11 (0.16)	0.12 (0.16)
Treat	0.19 (0.04)***	0.18 (0.04)***	0.18 (0.05)***	0.18 (0.05)***
Pretest	0.65 (0.02)***	0.61 (0.03)***	0.65 (0.02)***	0.62 (0.03)***
Treat x Pretest		0.08 (0.04)*		0.06 (0.04)
AIC	52345.52	52343.10	52310.11	52309.33
BIC	52388.92	52395.18	52370.87	52378.77
Log Likelihood	−26167.76	−26165.55	−26148.05	−26146.66
N observations	43480	43480	43480	43480
N students	2174	2174	2174	2174
N items	20	20	20	20
Var: Person	0.46	0.46	0.46	0.46
Var: Item	0.41	0.41	0.51	0.50
Var: Treatment			0.03	0.03
Cor: Item-Treatment			−0.74	−0.74

Notes: We apply the EIRM to data from Kim, et al. (2022), which examined the impact of a content literacy intervention. Column (1) presents a baseline constant effects model, column (2) presents the person-dependent HTE model, column (3) presents the item-dependent HTE model, and column (4) presents the flexible model allowing for both person- and item-HTE. The unit of observation is item-person for all the models. Standard errors are in parentheses. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Column 1 shows a clear positive ATE in log-odds units. That is, the treatment is estimated to cause a 0.19 logit increase in the probability of a correct response across all students and items, holding constant pretest scores. Adding the treatment by pretest interaction term in column 2 shows a small but significant positive interaction $\hat{\beta}_3 = 0.08$, suggesting that the treatment is more effective for previously high achieving students, in line with Figure 5. When we add ρ in column 3, we see that $\hat{\rho} = -.74$, indicating that treatment effects are most pronounced on the most difficult assessment items, as depicted in Figure 7. Finally, in the last column, we allow for both β_3 and ρ and see that the interaction term is rendered non-significant but $\hat{\rho}$ is unchanged. Likelihood ratio tests indicate that the specification in column 3 is preferred. Therefore, the observed HTE appears to be *not* due to the treatment

Figure 7: Correlation between item easiness and item treatment effect size



Notes: The points represent Empirical Bayes estimates of item easiness (b_i) and item-specific treatment effect size ($\beta_1 + \zeta_i$) derived from the item-dependent HTE model. The horizontal dashed lines represent the ATE and a null effect.

helping the highest achieving students most, but rather, impacting the hardest test items that better distinguish among previously high achieving students.

6 Discussion

Our study shows that using summary measures of student performance as outcome variables can lead to scenarios where two distinct data-generating processes produce virtually identical observed results. A treatment may appear more beneficial for students with high or low

baseline ability levels when effectiveness of the treatment genuinely varies based on students' baseline abilities (person-dependent HTE) or when the treatment impacts easier items more (or less) than hard ones (item-dependent HTE). These DGPs have distinct interpretations, and without item-level data, they cannot be distinguished. A treatment that disproportionately benefits low-scoring students across all items has different policy implications compared to one that uniformly enhances learning of easier content across all student groups. Consequently, standard analytic practices to estimate HTE may lead to misleading conclusions and poor policy decisions when incorrectly applied.

We show that this identification problem can be resolved by using an Explanatory Item Response Model (EIRM) that accounts for both variation in treatment effect along a pre-intervention participant characteristic and a correlation between item easiness and treatment effect size that can correctly determine the relevant DGP and draw correct conclusions. The EIRM is flexible and can be extended beyond the relatively simple application explored here to a wide array of data analytic settings, such as the inclusion of additional covariates at the person or item level, treatment by item cluster interactions, randomization blocks, multiple treatments, additional levels of hierarchy such as the nesting of students within schools, polytomous item responses, 2PL or 3PL models, or Bayesian approaches (Bulut et al., 2021; Bürkner, 2021; De Boeck et al., 2011; Gilbert, 2023b, 2024; Gilbert et al., 2023b; Stanke & Bulut, 2019), underscoring the applicability of our approach to diverse contexts. For researchers interested in applying our model to their own data sets, our references provide tutorials in the R programming language: the general EIRM for dichotomous items (De Boeck et al., 2011), extending the EIRM to polytomous items (Bulut et al., 2021), the item-dependent HTE model for dichotomous items (Gilbert, 2023b), and a Bayesian approach that allows for extensions including 2PL or 3PL models (Bürkner, 2021; Gilbert, 2023b). Furthermore, we provide both a detailed replication toolkit in online supplemental materials (OSM), and a brief appendix showing the basic R syntax to fit our model.

The problem of identification here hinges on item easiness varying in a systematic way such that $\rho \neq 0$. How realistic a problem is this likely to be in applied settings? This is a question that will merit more attention in future studies. One study of a 29-item reading comprehension assessment showed $\rho = .20$ (Gilbert et al., 2023b, p. 902). Similarly, an analysis of item-level data from 15 RCTs offers some evidence that $\rho \neq 0$ in that data (Ahmed et al., 2023, p. 8). Conceptually, the presence of DIF-by-easiness correlations in other settings (Duflo et al., 2011; Santelices & Wilson, 2010) is related and also suggestive of the possibility for such systematic variation. We largely leave this as an open question; whatever the value of ρ observed in any given setting, future research on HTE will be more sound if its value is scrutinized because it clearly plays an important role in subsequent inferences.

We used the example of pretest ability to motivate this study but our argument extends to *any* pre-treatment characteristic correlated with outcomes, such as age or socioeconomic status. In our OSM, we provide three additional empirical examples with alternative covariates (mathematics pretest, race indicator, high SES indicator) and an additional data set with item-level outcome data (a 36-item vocabulary assessment) to demonstrate this point. In all cases, ρ is strongly negative, and estimates of treatment by covariate interaction terms are attenuated when the model allows for ρ, findings fully consistent with our results showing bias in β_3 when $\rho \neq 0$. The generalizability of this phenomenon suggests that empirical researchers examining HTE may wish to adopt latent variable models, such as the EIRM, to avoid spurious conclusions about HTE. Furthermore, even if HTE is not of primary interest, $\rho \neq 0$ will necessarily create bias in *main* effects of covariates, because the main effect in a model without interactions an average of the two subgroup effects, weighted by the sample size of each subgroup.

While our exposition is primarily drawn from the perspective of education research, where analyses of student assessment data are ubiquitous, the implications of our study are relevant across various disciplines where both interest in HTE is high and detailed item-level data may potentially be available. Our results apply to economists, who collect item-level

data on consumption, spending, or attitudes (Jackson et al., 2021), psychologists studying traits such as depression through survey scales (Beevers et al., 2007; Gilbert et al., 2024; Schmitt et al., 2009), political scientists measuring political knowledge through questionnaires (Baek & Wojcieszak, 2009), or medical practitioners using surveys to supplement biometric measurement or clinical evaluation (B. W. Domingue et al., 2021; Hieronymus et al., 2019; Jessen et al., 2018). By making item-level data available, as advocated by B. Domingue and Kanopka, 2023, researchers can better analyze and interpret the impact of policies and interventions in education and other fields.

We acknowledge four limitations of our approach. First, item-level data is not always available to the researcher, which may preclude item-level analysis in many applications. Accordingly, it is unknown how prevalent or large item-easiness by treatment effect size correlations may be in empirical applications, and thus the scope of the problem outlined in this study is not well understood. Second, the logit coefficients produced by the EIRM are generally difficult to interpret compared to conventional test score metrics like standardized sum scores. While there are existing methods available to convert logit coefficients to standardized effect sizes (Breen et al., 2018; Gilbert et al., 2023b; Hox et al., 2017; Mood, 2010), these methods rest on strong assumptions and may add an extra layer of complexity for the researcher. For example, we can convert the main effect of the MORE treatment (Table 2, Model 1) to an effect size by the process of "y-standardization", in which we divide our coefficient β by the estimated pooled SD of θ_j^1, estimated as the residual standard deviation of the person random effect from a model with only the treatment indicator (Gilbert et al., 2023b, pp. 907-908, Footnote 4) (Briggs, 2008, p. 110). Applying this process to our data yields an estimated effect size of .20, which is likely to be more interpretable to practitioners than the logit coefficient. Third, the EIRM makes additional computational demands compared to alternative methods, due to the numerical integration required for parameter estimation (Rabe-Hesketh & Skrondal, 2022), particularly when employing Bayesian approaches requiring Markov chain Monte Carlo (MCMC) estimation (Bürkner, 2021; Gilbert, 2023b), and thus may

be time-consuming to employ on large data sets. Finally, we examined the relatively simple case of a randomized trial with two groups and a single time point, and as such it is unknown to what extent the issues identified here may generalize to alternative experimental and quasi-experimental contexts such as regression discontinuity, difference-in-differences, multisite trials, instrumental variables, and longitudinal analyses, though an emerging literature on the synthesis of latent variable and causal inference methods has begun to shed light on these areas (Gilbert et al., 2023a; Kuhfeld & Soland, 2022, 2023; Rabbitt, 2018; Soland, 2022, 2023; Soland et al., 2023).

Despite these limitations, our findings emphasize the critical role of measurement principles in program evaluation and demonstrate the useful implications of IRT-based causal analyses, beyond the traditional use of measurement tools solely for test construction. The adaptability of the EIRM makes it a powerful tool for any field that relies on detailed, item-level data to uncover patterns of treatment heterogeneity not readily apparent through more traditional data analysis methods.

Acknowledgements

This research was supported in part by the Jacobs Foundation. The authors wish to thank Zach Himmelsbach, Kenji Kitamura, Alex Bolves, Andrew Ho, the HGSE Measurement Lab, and the Stanford Psychometrics Lab for their helpful comments on drafts of this paper.

Data Availability

A replication toolkit containing our code and Online Supplemental Material is available via Research Box for researchers interested in replicating or extending the analyses in this study at the following URL: https://researchbox.org/2221. The code includes links to download the empirical data from Kim et al., 2022.

7 Appendix: R Code to fit the Explanatory Item Response Models used in this Study

The R code below illustrates how to fit various EIRMs to a data set `dat` with 0/1 outcome `correct`, 0/1 treatment indicator `treat`, baseline covariate `pretest`, person identifier `pid`, and item identifier `itemid`. For further resources, see the replication materials in our Online Supplemental Materials or the various EIRM R tutorials listed in the references.

```
# constant effects model
glmer(correct ~ treat + pretest + (1|pid) + (1|itemid),
      data = dat,
      family = binomial)

# person dependent HTE model
glmer(correct ~ treat*pretest + (1|pid) + (1|itemid),
      data = dat,
      family = binomial)

# item dependent HTE model
glmer(correct ~ treat + pretest + (1|pid) + (treat|itemid),
      data = dat,
      family = binomial)

# flexible model allowing for both person and item HTE
glmer(correct ~ treat*pretest + (1|pid) + (treat|itemid),
      data = dat,
      family = binomial)
```

References

Abenavoli, R. M. (2019). The mechanisms and moderators of "fade-out": Towards understanding why the skills of early childhood program participants converge over time with the skills of other children. *Psychological bulletin*, *145*(12), 1103.

Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., Rozelle, S., & Domingue, B. W. (2023). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Austin, P. C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Statistics in medicine*, *36*(20), 3257–3277.

Baek, Y. M., & Wojcieszak, M. E. (2009). Don't expect too much! learning from late-night comedy and knowledge item difficulty. *Communication Research*, *36*(6), 783–809.

Baker, F. B. (2001). *The basics of item response theory*. ERIC.

Beevers, C. G., Strong, D. R., Meyer, B., Pilkonis, P. A., & Miller, I. W. (2007). Efficiently assessing negative cognition in depression: An item response theory analysis of the dysfunctional attitude scale. *Psychological assessment*, *19*(2), 199.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Blundell, R., Dearden, L., & Sianesi, B. (2005). Evaluating the effect of education on earnings: Models, methods and results from the national child development survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(3), 473–512.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Brand, J. E., Pfeffer, F. T., & Goldrick-Rab, S. (2014). The community college effect revisited: The importance of attending to heterogeneity and complex counterfactuals.

Breen, R., Karlson, K. B., & Holm, A. (2018). Interpreting and understanding logits, probits, and other nonlinear probability models. *annual review of sociology*, *44*, 39–54.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89–118.

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour*, *5*(8), 980–989.

Bulut, O., Gorgun, G., & Yildirim-Erbasli, S. N. (2021). Estimating explanatory extensions of dichotomous and polytomous rasch models: The eirm package in r. *Psych*, *3*(3), 308–321.

Bürkner, P.-C. (2021). Bayesian item response modeling in r with brms and stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Camilli, G. (2006). Test fairness. *Educational measurement*, *4*, 221–256.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2022). *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india* (tech. rep.). National Bureau of Economic Research.

Colnet, B., Josse, J., Varoquaux, G., & Scornet, E. (2023). Risk ratio, odds ratio, risk difference... which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*.

De Boeck, P. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Science & Business Media.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software*, *39*, 1–28.

Ding, P., Feller, A., & Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, *114*(525), 304–317.

Domingue, B., & Kanopka, K. (2023). The item response warehouse (irw).

Domingue, B. W., Duncan, L., Harrati, A., & Belsky, D. W. (2021). Short-term mental health sequelae of bereavement predict long-term physical health decline in older adults: Us health and retirement study analysis. *The Journals of Gerontology: Series B*, *76*(6), 1231–1240.

Domingue, B. W., Kanopka, K., Trejo, S., Rhemtulla, M., & Tucker-Drob, E. M. (2022). Ubiquitous bias and false discovery due to model misspecification in analysis of statistical interactions: The role of the outcome's distribution and metric properties. *Psychological Methods*.

Domingue, B. W., Trejo, S., Armstrong-Carter, E., & Tucker-Drob, E. M. (2020). Interactions between polygenic scores and environments: Methodological and conceptual challenges. *Sociological Science*, *7*, 465–486.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American economic review*, *101*(5), 1739–1774.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378.

Francis, D. J., Kulesz, P. A., Khalaf, S., Walczak, M., & Vaughn, S. R. (2022). Is the treatment weak or the test insensitive: Interrogating item difficulties to elucidate the nature of reading intervention effects. *Learning and individual differences*, *97*, 102167.

Gilbert, J. B. (2023a). How measurement affects causal inference: Attenuation bias is (usually) more important than scoring weights. https://edworkingpapers.com/index.php/ai23-766

Gilbert, J. B. (2023b). Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in r. *Behavior Research Methods*.

Gilbert, J. B. (2024). Estimating treatment effects with the explanatory item response model. *Journal of Research on Educational Effectiveness*, *0*(0), 1–19. https://doi.org/10.1080/19345747.2023.2287601

Gilbert, J. B., Hieronymus, F., Eriksson, E., & Domingue, B. W. (2024). Item-level heterogeneous treatment effects of ssris on depression: Implications for inference, generalizability, and identification. *MedRxiv*.

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023a). Leveraging item parameter drift to assess transfer effects in vocabulary learning. (868). http://www.edworkingpapers.com/ai23-868

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023b). Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging large-scale online assessments to pinpoint the impact of educational interventions. *Journal of Educational and Behavioral Statistics*, *48*(6), 889–913.

Hieronymus, F., Lisinski, A., Nilsson, S., & Eriksson, E. (2019). Influence of baseline severity on the effects of ssris in depression: An item-based, patient-level post-hoc analysis. *The Lancet Psychiatry*, *6*(9), 745–752.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Jackson, K. C., Porte, S. C., Easton, J. Q., Blanchard, A., & Kiguel, S. (2021). Linking social-emotional learning to long-term success: Student survey responses show effects in high school and beyond. *Education Next*, *21*(1), 65–72.

Jeon, M., Jin, I. H., Schweinberger, M., & Baugh, S. (2021). Mapping unobserved item–respondent interactions: A latent space item response model with interaction map. *Psychometrika*, *86*(2), 378–403.

Jessen, A., Ho, A. D., Corrales, C. E., Yueh, B., & Shin, J. J. (2018). Improving measurement efficiency of the inner ear scale with item response theory. *Otolaryngology–Head and Neck Surgery*, *158*(6), 1093–1100.

Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & McIntyre, J. (2022). A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: Transfer effects on students' reading comprehension. *Journal of Educational Psychology*.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Teachers College Record*, *107*(14), 99–118.

Kuhfeld, M., & Soland, J. (2022). Avoiding bias from sum scores in growth estimates: An examination of irt-based approaches to scoring longitudinal survey responses. *Psychological Methods*, *27*(2), 234.

Kuhfeld, M., & Soland, J. (2023). Scoring assessments in multisite randomized control trials: Examining the sensitivity of treatment effect estimates to measurement choices. *Psychological Methods*.

Künzel, S. R., Walter, S. J., & Sekhon, J. S. (2019). Causaltoolbox—estimator stability for heterogeneous treatment effects. *Observational Studies*, *5*(2), 105–117.

Lockwood, J., & McCaffrey, D. F. (2020). Recommendations about estimating errors-in-variables regression in stata. *The Stata Journal*, *20*(1), 116–130.

Long, J. S. (1997). Regression models for categorical and limited dependent variables (vol. 7). *Advanced quantitative techniques in the social sciences*, *219*.

Lyu, W., Kim, J.-S., & Suk, Y. (2023). Estimating heterogeneous treatment effects within latent class multilevel models: A bayesian approach. *Journal of Educational and Behavioral Statistics*, *48*(1), 3–36.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior research methods*, *52*, 2287–2305.

Montoya, A. K., & Jeon, M. (2020). Mimic models for uniform and nonuniform dif as moderated mediation models. *Applied psychological measurement*, *44*(2), 118–136.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review*, *26*(1), 67–82.

Neuhaus, J. M., Kalbfleisch, J. D., & Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review/Revue Internationale de Statistique*, 25–35.

Olivera-Aguilar, M., & Rikoon, S. H. (2024). Intervention effect or measurement artifact? using invariance models to reveal response-shift bias in experimental studies. *Journal of Research on Educational Effectiveness*, *0*(0), 1–29. https://doi.org/10.1080/19345747.2023.2284768

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, *29*(3), 61–80.

Pearl, J. (2022). Detecting latent heterogeneity. In *Probabilistic and causal inference: The works of judea pearl* (pp. 483–506).

Petscher, Y., Compton, D. L., Steacy, L., & Kinnon, H. (2020). Past perspectives and new opportunities for the explanatory item response model. *Annals of dyslexia*, *70*, 160–179.

Rabbitt, M. P. (2018). Causal inference with latent variables from the rasch model as outcomes. *Measurement*, *120*, 193–205.

Rabe-Hesketh, S., & Skrondal, A. (2022). *Multilevel and longitudinal modeling using stata*. STATA press.

Reed, L. J., & Pearl, R. (1927). On the summation of logistic curves. *Journal of the Royal Statistical Society*, *90*(4), 729–746.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Sales, A., Prihar, E., Heffernan, N., & Pane, J. F. (2021). The effect of an intelligent tutor on performance on specific posttest problems. *International Educational Data Mining Society*.

San Martín, E. (2016). Identification of item response theory models. *Handbook of item response theory*, *2*, 127–150.

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? the case of freedle, the sat, and the standardization approach to differential item functioning. *Harvard Educational Review*, *80*(1), 106–134.

Schmitt, A. B., Bauer, M., Volz, H.-P., Moeller, H.-J., Jiang, Q., Ninan, P. T., & Loeschmann, P.-A. (2009). Differential effects of venlafaxine in the treatment of major depressive disorder according to baseline severity. *European archives of psychiatry and clinical neuroscience*, *259*, 329–339.

Schochet, P. Z., Puma, M., Deke, J., et al. (2014). Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods. *US Department of Education, Washington, DC. Report No. NCEE*, *4017*.

Schuetze, B. A., & von Hippel, P. T. (2023). How not to fool ourselves about heterogeneity of treatment effects. *PsyArXiv. October*, *23*.

Sijtsma, K. (2005). Nonparametric item response theory models. *Encyclopedia of social measurement*, *2*, 875–882.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*, 563–575.

Soland, J. (2022). Evidence that selecting an appropriate item response theory–based approach to scoring surveys can help avoid biased treatment effect estimates. *Educational and Psychological Measurement*, *82*(2), 376–403.

Soland, J. (2023). Item response theory models for difference-in-difference estimates (and whether they are worth the trouble). *Journal of Research on Educational Effectiveness*, 1–31.

Soland, J., Johnson, A., & Talbert, E. (2023). Regression discontinuity designs in a latent variable framework. *Psychological Methods*, *28*(3), 691.

Soland, J., Kuhfeld, M., & Edwards, K. (2022). How survey scoring decisions can influence your study's results: A trip through the irt looking glass. *Psychological Methods*.

Spoto, A., & Stefanutti, L. (2023). Empirical indistinguishability: From the knowledge structure to the skills. *British Journal of Mathematical and Statistical Psychology*, *76*(2), 312–326.

Stanke, L., & Bulut, O. (2019). Explanatory item response models for polytomous item responses. *International Journal of Assessment Tools in Education*, *6*(2), 259–278.

Torche, F., Fletcher, J., & Brand, J. E. (2024). Disparate effects of disruptive events on children. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *10*(1), 1–30.

VanderWeele, T. J., & Batty, C. J. (2023). On the dimensional indeterminacy of one-wave factor analysis under causal effects. *Journal of Causal Inference*, *11*(1), 20220074.

VanderWeele, T. J., & Vansteelandt, S. (2022). A statistical test to reject the structural interpretation of a latent factor model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(5), 2032–2054.

Von Hippel, P. (2015). Linear vs. logistic probability models: Which is better, and when. *Statistical Horizons*.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wallace, M. L., Mentch, L., Wheeler, B. J., Tapia, A. L., Richards, M., Zhou, S., Yi, L., Redline, S., & Buysse, D. J. (2023). Use and misuse of random forest variable importance metrics in medicine: Demonstrations through incident stroke prediction. *BMC Medical Research Methodology*, *23*(1), 144.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*, *37*(23), 3309–3324.

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*(2), 788–806.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual review of sociology*, *25*(1), 659–706.

Wolf, B., & Harbatkin, E. (2023). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, *16*(1), 134–161.

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, *42*(1), 314–347.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., et al. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369.

Zisook, S., Shear, K., & Kendler, K. S. (2007). Validity of the bereavement exclusion criterion for the diagnosis of major depressive episode. *World Psychiatry*, *6*(2), 102.

8 Authors

Joshua B. Gilbert is a PhD student in the Education Policy and Program Evaluation Concentration at Harvard University Graduate School of Education. His research interests include the intersection of psychometric and causal inference methods; e-mail: joshua_gilbert@g.harvard.edu.

Luke W. Miratrix is an associate professor at the Harvard Graduate School of Education. His research interests are primarily pertaining to causality with a focus on developing methodology to assess and characterize treatment effect heterogeneity in randomized clinical trials and observational studies; e-mail: luke_miratrix@gse.harvard.edu.

Mridul Joshi is a PhD student in Economics and Education at Stanford University. His research focuses on the economics of education and development economics; e-mail: mriduljoshi@stanford.edu

Benjamin W. Domingue is an associate professor at the Stanford Graduate School of Education. He is interested in psychometrics and quantitative methods; e-mail: bdomingue@stanford.edu