ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-619

Modeling Item-Level Heterogeneous Treatment Effects with the Explanatory Item Response Model: Leveraging Online Formative Assessments to Pinpoint the Impact of Educational

Joshua B. Gilbert
Harvard University

James S. Kim
Harvard University

Luke W. Miratrix
Harvard University

Analyses that reveal how treatment effects vary allow researchers, practitioners, and policymakers to better understand the efficacy of educational interventions. In practice, however, standard statistical methods for addressing Heterogeneous Treatment Effects (HTE) fail to address the HTE that may exist within outcome measures. In this study, we present a novel application of the Explanatory Item Response Model (EIRM) for assessing what we term "item-level" HTE (IL-HTE), in which a unique treatment effect is estimated for each item in an assessment. Results from data simulation reveal that when IL-HTE are present but ignored in the model, standard errors can be underestimated and false positive rates can increase. We then apply the EIRM to assess the impact of a literacy intervention focused on promoting transfer in reading comprehension on a digital formative assessment delivered online to approximately 8,000 third-grade students. We demonstrate that allowing for IL-HTE can reveal treatment effects at the item-level masked by a null average treatment effect, and the EIRM can thus provide fine-grained information for researchers and policymakers on the potentially heterogeneous causal effects of educational interventions.

VERSION: August 2022

Suggested citation: Gilbert, Joshua B., James S. Kim, and Luke W. Miratrix. (2022). Modeling Item-Level Heterogeneous Treatment Effects with the Explanatory Item Response Model: Leveraging Online Formative Assessments to Pinpoint the Impact of Educational Interventions. (EdWorkingPaper: 22-619). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/m3jh-kh96

Modeling Item-Level Heterogeneous Treatment Effects with the Explanatory Item Response Model: Leveraging Online Formative Assessments to Pinpoint the Impact of Educational Interventions

Joshua B. Gilbert, James S. Kim, and Luke W. Miratrix

Harvard University Graduate School of Education

Corresponding Author: joshua_gilbert@g.harvard.edu

Abstract

Analyses that reveal how treatment effects vary allow researchers, practitioners, and policymakers to better understand the efficacy of educational interventions. In practice, however, standard statistical methods for addressing Heterogeneous Treatment Effects (HTE) fail to address the HTE that may exist *within* outcome measures. In this study, we present a novel application of the Explanatory Item Response Model (EIRM) for assessing what we term "item-level" HTE (IL-HTE), in which a unique treatment effect is estimated for each item in an assessment. Results from data simulation reveal that when IL-HTE are present but ignored in the model, standard errors can be underestimated and false positive rates can increase. We then apply the EIRM to assess the impact of a literacy intervention focused on promoting transfer in reading comprehension on a digital formative assessment delivered online to approximately 8,000 third-grade students. We demonstrate that allowing for IL-HTE can reveal treatment effects at the item-level masked by a null average treatment effect, and the EIRM can thus provide fine-grained information for researchers and policymakers on the potentially heterogeneous causal effects of educational interventions.

Keywords: Heterogeneous Treatment Effects, Explanatory Item Response Model, Causal Inference, Simulation, Psychometrics

Modeling Item-Level Heterogeneous Treatment Effects with the Explanatory Item Response Model: Leveraging Online Formative Assessments to Pinpoint the Impact of Educational Interventions

Analyses that explore Heterogeneous Treatment Effects (HTE) are increasingly becoming standard in education research, as understanding how and why treatment effects vary is critical for the translation of academic research to the implementation of educational interventions (Schochet, Puma, & Deke, p. 1). Traditional methodological approaches to HTE such as subgroup analysis, moderation (i.e., statistical interaction), reweighting for generalization, mediation, instrumental variables estimation, and quantile regression all provide critical insight into the potentially varying impacts of an educational intervention, but ignore the most fine-grained perspective on how treatment effects may vary *within* an outcome measure itself. In this study, we aim to expand the analyst's HTE toolkit by proposing and testing a novel application of the Explanatory Item Response Model (EIRM; De Boeck & Wilson, 2004) for assessing what we term "item-level" Heterogeneous Treatment Effects (IL-HTE). That is, treatment effects may differ not just between demographic subgroups or according to some baseline characteristic such as pretest scores, as in traditional HTE analysis, but across the various items of an outcome measure, such as an educational assessment, manifested by treatment effects that vary at the item level. This methodological gap can be addressed with the EIRM because it models individual item responses directly rather than as a single summary value such as a sum score or IRT-based ability estimate, thereby allowing researchers to assess the presence of IL-HTE and to quantify its explained and unexplained sources.

The EIRM has been applied primarily to psychometric research questions such as the relationship between person or item characteristics and item response patterns (see for example

Kim, et al., 2010, which uses an EIRM to assess the predictors of letter-sound acquisition in an observational study). However, the EIRM has seen less application in causal inference contexts despite its theoretical appeal and its ability to combine measurement (i.e., psychometric) and explanatory (i.e., regression) models into a single computational procedure (Briggs, 2008; Christensen, 2006; Rabbitt, 2018; Zwinderman, 1991), and we are aware of no methodological or empirical studies to date that employ the EIRM to explore IL-HTE. By explicitly modeling IL-HTE using the novel approach presented in this study, and in some cases, uncovering statistically significant item-level treatment effects masked by a null average treatment effect, the EIRM allows researchers to gain more fine-grained insight into the efficacy of educational interventions. This fine-grained insight in turn allows researchers to contextualize and interpret impact analyses in ways that are more actionable for practitioners and policymakers, and ultimately supports the goal of more targeted diagnosis and intervention to support student learning outcomes.

This study introduces a general approach for conducting IL-HTE analysis within the context of a large-scale, cluster-randomized controlled trial that involved third grade students from every K-5 elementary school in one of the largest school districts in the United States (k = 110 schools, n = 7797 students). The RCT tests the efficacy of the Model of Reading Engagement (MORE) intervention, which emphasizes thematic lessons that provide an intellectual framework for building domain knowledge to help third-grade students connect new learning to a general schema and to transfer their learning to novel reading comprehension tasks (see Kim, et al., 2021; Kim, et al., 2022 for a detailed description of MORE and prior research results). In the MORE intervention, the general schema for the concept of *system* (i.e., how systems function properly) was induced through a 12-day science lesson sequence focused on the topic of human body systems. All schools implemented the 12-day lesson sequence on human body systems and were

randomly assigned to implement two additional lessons that involved either a double dose of science vocabulary and concepts through a read aloud text on the human body system and stem cells (control), or social studies extension lessons on collaborative systems focused on how leaders collaborated in the Apollo 11 moon mission (treatment). That is, the RCT aimed to test the hypothesis that students could leverage the general schema for *system* through repeated exposure to a science topic (i.e., human body systems) and brief exposure to social studies topic (i.e., collaborative systems) while reading unfamiliar science and social studies passages to demonstrate learning on an online formative reading comprehension assessment.

The formative assessment included three reading comprehension transfer tasks that varied by passage-item type (Near, Mid, and Far transfer passages determined by the number of science and social studies words that appeared in the lesson texts) and was administered electronically to all third graders in the study. Following the intervention implementation, we provided superintendents, principals, and teachers with detailed item- and passage-level information from the assessment. Here, we extend the descriptive analyses provided to participants by statistically evaluating IL-HTE to assess potential transfer effects on reading comprehension, thus illustrating how a novel application of the EIRM can provide immediate, fine-grained, population-level evidence of causal impact and can potentially help decision-makers diagnose and intervene to support students before the administration of the end-of-grade three reading test, used for high-stakes accountability purposes (i.e., threat of grade retention and required summer school). The full assessment is available in the Online Supplemental Materials.

Methodologically, we pursue two aims. First, a data simulation to assess the performance of the EIRM in the presence of IL-HTE and the related conceptual issues that arise, and second, an application of the EIRM to empirical educational assessment data from the MORE intervention.

A replication toolkit is available from the authors for researchers interested in replicating or extending the simulation or the analysis of the assessment data.

The Explanatory Item Response Model (EIRM)

Because the statistical theory underlying the EIRM has been described extensively in prior literature, we provide only a brief review here. Readers interested in further details about the EIRM are directed to Wilson, De Boeck, and Carstensen (2008) for a short introduction, De Boeck, Cho, and Wilson (2016) for a recent review, and De Boeck and Wilson, (2004) for a book-length treatment. For a detailed review of generalized linear mixed models (GLMMs), of which the EIRM is a special case, see Stroup (2012). For a practical introduction to fitting the EIRM in R with the `lme4` package, see De Boeck, et al., (2011).

The EIRM is a cross-classified multilevel logistic regression model, in which item responses are nested within the cross-classification of persons and items. In its simplest form with random effects for persons and items, it can be expressed as

$$logit\left(P\left(y_{ij}=1\right)\right)=\theta_j+\zeta_i$$

$$\theta_j \sim N\left(0,\sigma_\theta^2\right)$$

$$\zeta_i \sim N(0,\sigma_{\zeta_0}^2)$$

in which the log-odds of a correct response to item i for person j is a function of person ability θ_j and item easiness ζ_i (item easiness is the negative of what is often called item difficulty in the Item Response Theory literature). The EIRM with no person or item predictors is equivalent to the Rasch or One-Parameter Logistic (1PL) IRT model when the item easiness parameters are considered fixed.

An important modelling choice when employing the EIRM is the distinction between fixed and random effects for items and persons. In the IRT and EIRM contexts, persons are almost

always modeled as random effects, that is, as normally distributed with mean zero and an unknown variance, but we have a choice between fixed and random effects for the assessment items (De Boeck, 2008). Random effects allow for estimation of the distributions of item easiness or student abilities. When referencing, for example, item easiness against the standard deviation of student ability, we can better understand the range of difficulties of the items on the test.

In base form, the EIRM with random person and item effects is commonly called a "doubly descriptive" model (Wilson, De Boeck, & Carstensen, 2008, p. 95) as it solely provides estimates of the variances of both persons and items without any variables to *explain* systematic differences in person ability or item easiness. The EIRM becomes "person explanatory" or "item explanatory" when predictors at the person or item level are added to the model, or "doubly explanatory" when both person and item level predictors are included. As such, the EIRM can address research questions at the person level (e.g., do older students have systematically higher probabilities of a correct response) or at the item level (e.g., are items that assess phonological awareness systematically more difficulty than items that assess vocabulary), or both (e.g., do male-female performance gaps depend on item type). While the EIRM has primarily been applied to observational studies to examine person- or item-level predictors of response patterns, it can easily be applied to causal inference contexts, and ultimately to the possibility of examining IL-HTE, by including a person-level treatment variable in the model, a possibility to which we now turn.

Modeling Item Level Heterogeneous Treatment Effects

We can model IL-HTE by introducing an interaction between item and treatment assignment in an EIRM through a random slope term. To illustrate, consider the following two models, presented in reduced form:

Model 1 – Constant Treatment Effect:

$$logit\left(P(y_{ij}=1)\right)=\beta_0+\beta_1 treat_j+\theta_j+\zeta_{0i}$$

$$\theta_j \sim N(0,\sigma_\theta^2)$$

$$\zeta_{0i} \sim N(0,\sigma_{\zeta_0}^2)$$

Model 2 – IL-HTE:

$$logit\left(P(y_{ij}=1)\right)=\beta_0+\beta_1 treat_j+\theta_j+\zeta_{0i}+\zeta_{1i} treat_j$$

$$\theta_j \sim N(0,\sigma_\theta^2)$$

$$\begin{bmatrix}\zeta_{0i}\\ \zeta_{1i}\end{bmatrix} \sim N\left(0,\begin{bmatrix}\sigma_{\zeta_0}^2 & \rho_{10}\\ \rho_{01} & \sigma_{\zeta_1}^2\end{bmatrix}\right),$$

in which y_{ij} is the dichotomous response to item i for person j, β_0 is the log-odds of a correct response for a student of average ability to an item of average difficulty, β_1 is the average treatment effect (ATE) across items, θ_j is a random intercept representing unexplained person ability (as in traditional IRT modeling), and ζ_{0i} is a random intercept representing item easiness (equivalent to the negative of the item difficulty parameter in an IRT context).

The difference between Model 1 and Model 2 is the random slope, ζ_{1i}, that captures the deviation between each item's individual treatment effect and the ATE β_1, thus allowing for IL-HTE. Model 2 also allows for correlation between item easiness and item treatment effect size (ρ_{01}). We can additionally include item-level characteristics interacted with treatment to assess systematic treatment variation; the random slopes represent idiosyncratic, or unexplained, variation (Ding, Feller, & Miratrix, 2019). In other words, paraphrasing Raudenbush and Bloom (2015), the random slopes allow us to learn *about* the presence of IL-HTE, and treatment by item interactions allow us to learn *from* IL-HTE.

While IL-HTE could in principle be modeled with the combination of item fixed effects and treatment by item interaction terms, the fully fixed effects approach is suboptimal for our

purposes for several reasons. First, the effects of item characteristics are not estimable when fixed item effects are used because, as item-level covariates, they would be collinear with the item indicators. Second, for a fully fixed model, an additional treatment-by-item interaction term would be needed for each item, adding complexity to the model, whereas the random effects model includes a single variance component for the treatment effect (i.e., the random slope) and is therefore more parsimonious. (One could use a fixed-intercept, random slope formation where the item effects were fixed but the interaction terms were random, as described in Bloom et al. (2017); we do not study this possibility here.) Third, the random effects approach provides a direct parameter estimate of the degree of IL-HTE present in the data through variance of the treatment coefficient, a parameter of interest that has no analogue in fixed effects analysis. Fourth, shrinkage provides more stable estimates of the individual item difficulties and item-level treatment effects, an especially important benefit unless dataset sizes are very large. Last, and most important for our purposes, the random effects parameterization better matches our focus on IL-HTE as it explicitly models items as a source of variability due to taking the test items as being (possibly literally) drawn from a pool of potential items.

That is, at a conceptual level, an item fixed effect model does not take the variability of which items are included on a test into account, and therefore the associated uncertainty estimates will be relative to the test-specific estimand of the true ATE across the items in the realized test, rather than across the (possibly hypothetical) population of items that *could have been* on the test. In other words, when IL-HTE is present, a given draw of items will have its own finite sample ATE that differs from that of the population of items due to sampling error. For example, if the test happens to include items that are more sensitive to the treatment than other items that might have been included, the test-specific estimand would be larger, the point estimate of average

treatment impact would tend to be larger, and the fixed-effect estimated standard errors would reflect estimation uncertainty relative to the test-specific estimand, *not* the population average estimand. In contrast, the random slope model that allows for IL-HTE would target the mean treatment effect in the *population* of items from which a test is (hypothetically) constructed, and the associated uncertainty estimates would incorporate the additional uncertainty of which items are selected for a test administration. The contrast between finite sample and population average estimands in the EIRM is analogous to fixed and random effect estimators for ATEs in multisite trials (Miratrix, et al., 2021, p. 280; Chan & Hedges, 2022) or meta-analysis (Skrondal & Rabe-Hesketh, 2004, Chapter 9).

Importantly, the constant effect model, with item random intercepts but no random slopes, directly corresponds to the item fixed effect model. In fact, as shown in Miratrix et al. (2021), the constant effect model estimates a precision-weighted estimand of the item-level average treatment effects, but so long as each student takes the same test, and all items have equal numbers of observations, the precision-weighted point estimate of the ATE will exactly coincide with that of the fixed-effect model. In other words, ignoring IL-HTE provides inference for the test-specific ATE, and ignores any additional uncertainty due to whether the selected test items are representative. If there is substantial IL-HTE, ignoring such uncertainty could be misleading as we generally are interested in the underlying construct being measured, not whether treatment happened to impact students as measured by the specific items selected. Consider, for example, that if researchers could somehow *a priori* select those items known to be more sensitive to the treatment, they would obtain a larger measured treatment impact as an artifact of the selected items, rather than a truly more effective treatment.

Overall, we argue that item random effects with a randomly varying treatment coefficient is generally the more appropriate choice for modelling IL-HTE. A fixed effect or constant effect model would be preferred when only the finite sample ATE across the specific items of the administered test is of interest, such as when the assessment has a fixed set of items across replications, and these items are viewed as fully encompassing the scope of what is being measured.

Monte Carlo Simulation

To illustrate the ability of the EIRM to account for IL-HTE, we first conduct a simulation comparing our two base modeling approaches across a range of contexts. We generate data from our IL-HTE model with normally distributed error terms and no correlation between item difficultly and item-level treatment impact. We fixed the number of subjects at 500 and the number of items at 20 and explored the combination of two varying simulation factors: (1) the average treatment effect size on the logit scale (0 and 0.4) and (2) the standard deviation of item-level treatment effects (0 for no HTE, 0.2 for moderate HTE, and 0.4 for high HTE). Thus, we employed a 2×3 full factorial design examining null and positive average treatment effect sizes fully crossed with no, moderate, and high IL-HTE for a total of six parameterizations. Each parameterization was replicated 2000 times for a total of 12,000 simulated data sets, in which we generate a new set of 20 items according to our parameters and then simulate our experiment using those 20 items as our test. We then fit our two EIRMs as cross-classified logistic regression models (i.e., generalized linear mixed models with a logit link function and random effects for students and items) using the `glmer` function from the R package `lme4` (Bates, et al., 2015) to estimate the model parameters for each simulated data set, one constraining the treatment effect to be constant (i.e., no IL-HTE), the other allowing for IL-HTE, and collected the model output for further analysis.

Empirical Assessment Data

For our empirical application, we examine the intention-to-treat (ITT) impact of the Model of Reading Engagement (MORE) intervention on third grade reading comprehension from a cluster-randomized controlled trial. Our data, collected in the 2021-2022 school year, consists of 110 schools randomly assigned to treatment and control from a large urban district in the southeastern United States (N = 7797 students). We examine dichotomous (correct/incorrect) student responses on a researcher-designed reading comprehension assessment containing 30 multiple-choice items based on three reading passages designed to measure different degrees of transfer from the MORE intervention curriculum (i.e., Near, Mid, and Far Transfer passages, with varying numbers of words from the MORE lessons). The formative assessment was administered online at the end of the MORE intervention, but prior to the high-stakes end-of-year state test.

The primary substantive research aim was to understand whether students could leverage the general schema for *system* in comprehending novel passages related to social studies topics after learning about various human body systems. Thus, we hypothesized that control students (who received a double dose of science lessons) and treatment students (who received two social studies extension lessons) would perform equally well on the Near Transfer items with only science concepts. Furthermore, if treatment students could successfully leverage their vocabulary knowledge from the social studies extension lessons while reading the Mid and Far Transfer passages, we hypothesized that treatment students would outperform control students on Mid Transfer items and potentially also on Far Transfer items.

As such, we fit four EIRMs to the data, modeling the probability of correct response to item i for student j in school k, presented below in reduced form:

Model 1 – MORE Assessment EIRM 1, No IL-HTE:

$$logit\left(P\left(y_{ijk}=1\right)\right)=\beta_0+\beta_1 treat_k+\beta_2 pretest_{jk}+\ \theta_{jk}+\zeta_{0i}+\nu_k$$

$$\theta_{jk}\sim N(0,\sigma_\theta^2)$$

$$\zeta_{0i}\sim N(0,\sigma_{\zeta_0}^2)$$

$$\nu_k\sim N(0,\sigma_\nu^2)$$

Model 2 – MORE Assessment EIRM 2, Randomly Varying IL-HTE:

$$logit\left(P\left(y_{ijk}=1\right)\right)=\beta_0+\beta_1 treat_k+\beta_2 pretest_{jk}+\ \theta_{jk}+\zeta_{0i}+\zeta_{1i}treat_k+\nu_k$$

$$\theta_{jk}\sim N(0,\sigma_\theta^2)$$

$$\begin{bmatrix}\zeta_{0i}\\ \zeta_{1i}\end{bmatrix}\sim N\left(0,\begin{bmatrix}\sigma_{\zeta_0}^2 & \rho_{10}\\ \rho_{01} & \sigma_{\zeta_1}^2\end{bmatrix}\right)$$

$$\nu_k\sim N(0,\sigma_\nu^2)$$

Model 3 – MORE Assessment EIRM 3, Systematically and Randomly Varying IL-HTE:

$$\begin{aligned}&logit\left(P\left(y_{ijk}=1\right)\right)\\&\quad=\beta_0+\beta_1 treat_k+\beta_2 pretest_{jk}+\beta_3 mid_i+\beta_4 far_i+\beta_5 treat\times mid_i\\&\quad+\ \beta_6 treat\times far_i+\ \theta_{jk}+\zeta_{0i}+\zeta_{1i}treat_k+\nu_k\end{aligned}$$

$$\theta_{jk}\sim N(0,\sigma_\theta^2)$$

$$\begin{bmatrix}\zeta_{0i}\\ \zeta_{1i}\end{bmatrix}\sim N\left(0,\begin{bmatrix}\sigma_{\zeta_0}^2 & \rho_{10}\\ \rho_{01} & \sigma_{\zeta_1}^2\end{bmatrix}\right)$$

$$\nu_0\sim N(0,\sigma_\nu^2)$$

Model 4 – MORE Assessment EIRM 4, Systematically Varying IL-HTE:

$$\begin{aligned}&logit\left(P\left(y_{ijk}=1\right)\right)\\&\quad=\beta_0+\beta_1 treat_k+\beta_2 pretest_{jk}+\beta_3 mid_i+\beta_4 far_i+\beta_5 treat\times mid_i\\&\quad+\ \beta_6 treat\times far_i+\ \theta_{jk}+\zeta_{0i}+\nu_k\end{aligned}$$

$$\theta_{jk} \sim N(0, \sigma_{\theta}^2)$$

$$\zeta_{0i} \sim N(0, \sigma_{\zeta_1}^2)$$

$$\nu_k \sim N(0, \sigma_{\nu}^2).$$

All EIRM parameters are interpreted analogously to those of the simulation models described earlier, with addition of the subscript k indexing school membership, a random intercept for school (ν_k) to account for the cluster-randomized design, a student-level reading pretest score to improve the precision of the estimates (β_2), main effects for passage easiness (β_3, β_4), and passage by treatment interactions (β_5, β_6) to model systematic sources of IL-HTE. Across all EIRMs, we assess the statistical significance of the fixed effects via Wald tests (for individual coefficients) and likelihood ratio tests (for sets of coefficients such as the passage by treatment interactions), and that of the random effects by likelihood ratio tests comparing nested models with and without the random effects of interest.

Results

Simulation

The results of the simulation reveal first that the point estimates for average treatment effects for the constant and IL-HTE models are nearly identical (r = 0.998). Analysis of the uncertainty associated with the point estimates is more complex. The top panel of Figure 1 compares the mean of the estimated standard errors (SEs) of the average treatment effect point estimates to their true standard errors (i.e., the observed standard deviation of the treatment effect point estimates) in a scatterplot, in which we would expect the points to fall on the $y = x$ identity line if the model is well calibrated. We see that the model-based SEs are well calibrated for the HTE model, but the model-based SEs for the constant treatment effect model are systematically too low when IL-HTE is high, falling below the diagonal identity line. The constant treatment

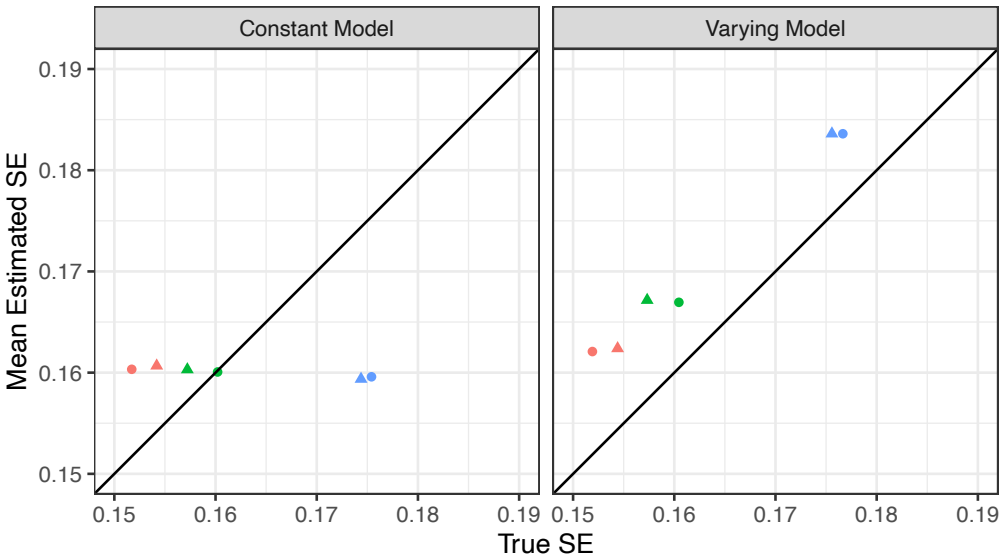
effect model, like a fixed effect model, is not accounting for the additional uncertainty of whether the selected test items are representative of the full item bank.

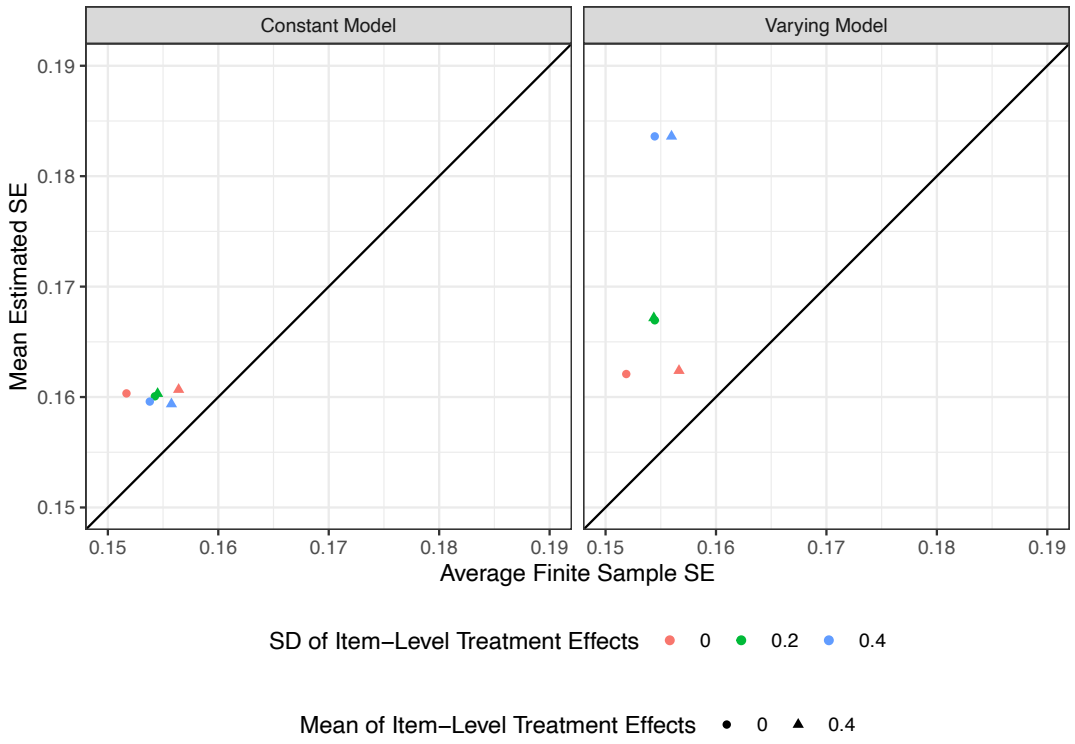
However, when we instead compare the average estimated SEs to the standard deviation of the *finite sample* ATEs (equivalent to the true finite-sample SEs averaged across the different sets of simulated test items), as shown in the bottom panel of Figure 1, we clearly observe that the estimated SE of the constant treatment effect EIRM is better calibrated, regardless of the level of IL-HTE. Therefore, the choice to allow IL-HTE in an EIRM is not just a statistical issue, but a substantive one, and researchers should consider what estimand they intend to target when selecting a modeling strategy.

Figure 1. Comparison of Estimated and True SEs of EIRMs with and without IL-HTE

Top: Item population estimand (IL-HTE)

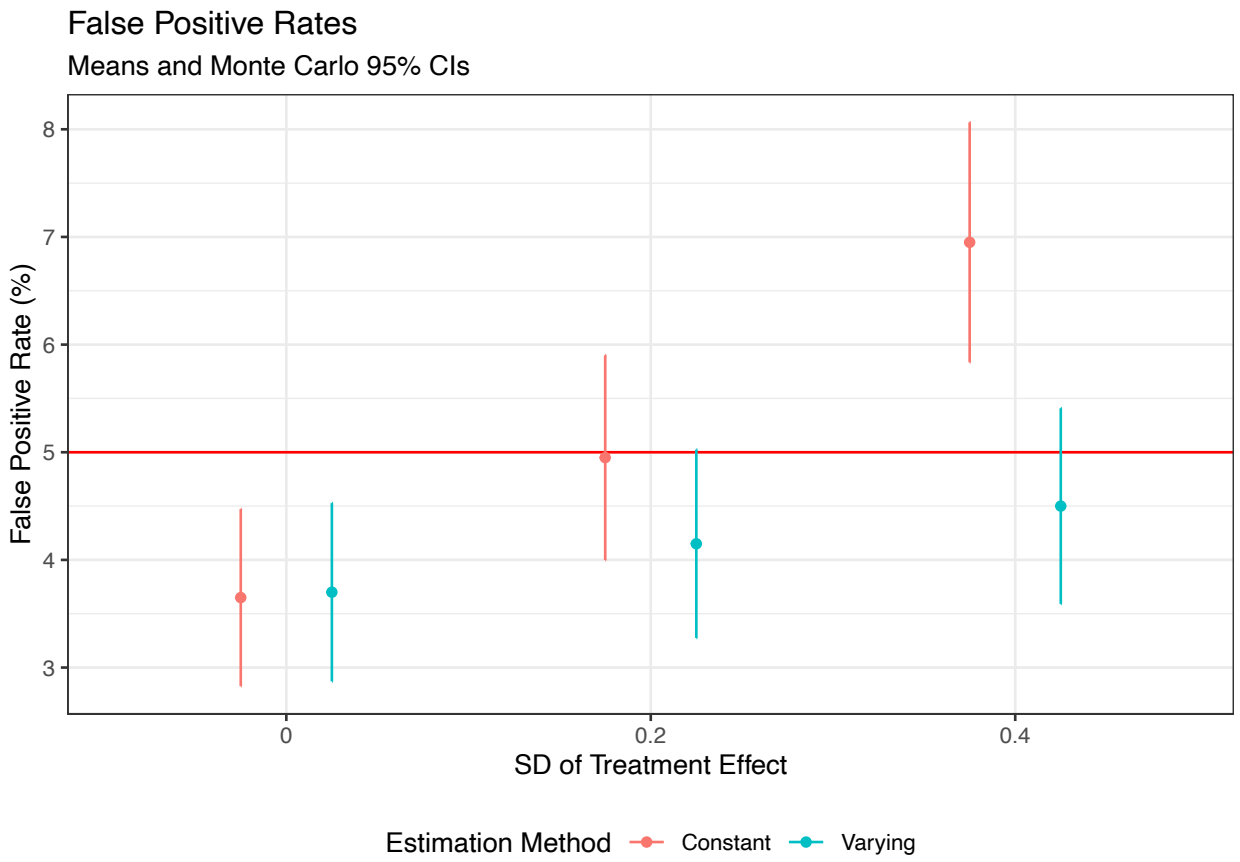
Bottom: Test-specific estimand (No IL-HTE)





Proceeding under the assumption that the ATE in the population of items is the estimand of interest, the practical effect of ignoring IL-HTE when it is present is depicted in Figure 2, which provides the estimated false positive rates for each estimation method at each level of IL-HTE. The false positive rate increases for the constant effect EIRM as IL-HTE rises, whereas the false positive rates are close to the nominal value of 5% when the treatment effect is allowed to vary at the item level, indicating that ignoring potential IL-HTE provides unrealistically precise estimates of average treatment effects, with systematically underestimated SEs and invalid hypothesis tests. These findings are consistent with prior simulation studies on the importance of including random coefficients in mixed-effects models more generally (Bell, Fairbrother, & Jones, 2019, pp. 1062-1065).

Figure 2: Comparison of false positive rates by method based on true IL-HTE



Application to MORE Empirical Assessment Data

The results of the four EIRMs applied to the MORE intervention data are summarized in Table 1. Model 1 shows that the average MORE treatment effect across all reading comprehension items is positive but not statistically significant (β_1 = 0.05, p = 0.53). Without considering the possibility of IL-HTE, an analyst might stop at this step and conclude that there is no effect of the MORE intervention on student reading comprehension. However, Model 2 shows statistically significant and substantively meaningful item-level treatment effect variation, as the SD of the randomly varying item-level treatment effect (σ_{ζ_1} = 0.05, p < 0.05) is as large as the point estimate itself (β_1 = 0.05) and implies a 95% prediction interval of approximately -0.05 to +0.15 for individual item-level treatment effects on the logit scale. Model 3 tests the hypothesis that treatment effects systematically depend on the passage type by including the treatment by passage

type interaction terms. Results show that while the treatment effects on the Near and Far Transfer reading passages are not distinguishable from zero, items from the Mid Transfer passage show a significantly larger average treatment effect than the other passage types ($\beta_5 = 0.07$, $p < 0.05$). These results are consistent with the hypothesis that there would be no treatment-control differences on the Near Transfer passage but rather that the social studies extension lessons provided to treatment students had a positive effect on the Mid Transfer passage items, which included science and social studies vocabulary words. We can also assess whether item difficulty is associated with sensitivity to treatment. In Model 3, for example, the estimated item easiness and treatment effect correlation is large in magnitude ($\rho_{12} = 0.55$, $p = 0.12$), suggesting that within each passage, the easier items are most responsive to the MORE intervention, though the correlation is non-significant due to the small number of items per passage.

Finally, and with the caveats about cross-model comparisons of variance components in the logistic context in mind (Hox, Moerbeek, & Van de Schoot, 2017, pp. 121-128), we see that the treatment by passage type interaction effect explains roughly 64% $\left(\frac{0.05^2-0.03^2}{0.05^2}\right)$ of the treatment effect variance. Accordingly, Model 4 tests the hypothesis that the treatment by passage type interaction explains all IL-HTE by removing the random slope for treatment from the model. A likelihood ratio test shows that Model 3, which includes the random slope, is not distinguishable from Model 4, which omits the random slope, and therefore we can conclude that the treatment by passage type interaction could be capturing all IL-HTE in this data set. Substantively, the MORE intervention appears to have its strongest impact on Mid Transfer items, suggesting that even a brief exposure to targeted social studies vocabulary words and concepts through read-aloud lessons can lead to measurable improvements in reading comprehension that involves those same vocabulary words and concepts, particularly when students can access and extend a general

schema for *system* that was taught in previous science and social studies lessons. Such a fine-grained understanding of the efficacy of the MORE intervention on reading comprehension would have been ignored had we not considered the possibility of IL-HTE, or instead had examined a single classical test theory or IRT-based summary score.

A visualization of the randomly varying item-level treatment effects of Model 2 are displayed in the top panel of Figure 3, in which the dashed red line shows the average treatment effect, and the points show item-specific treatment effects and 95% CIs on the logit scale and are color coded by passage type. Forecasting the results of Models 3 and 4, we can see that the Mid Transfer item-level treatment effects (green points) are concentrated on the high end of the treatment effect distribution. The bottom panel shows the population average probabilities of a correct response as a function of subtest passage type and treatment status based on Model 4, confirming that the average treatment-control difference is largest on the Mid Transfer passage on the probability scale.

Table 1. Results of Explanatory Item Response Models fitted to the MORE Reading Comprehension Assessment Data

Parameter	Model 1: No IL-HTE	Model 2: Randomly Varying IL-HTE	Model 3: Systematically and Randomly Varying IL-HTE	Model 4: Systematically Varying IL-HTE
Fixed Effects				
Treatment (β_1)	0.05 (0.08)	0.05 (0.08)	0.03 (0.08)	0.02 (0.08)
Treatment x Mid (β_5)			0.07 (0.03)**	0.08 (0.02)**
Treatment x Far (β_6)			0 (0.03)	0 (0.02)
Variance Components (SDs and Correlations)				
Student (σ_θ)	0.63	0.63	0.63	0.63
School (σ_ν)	0.39	0.39	0.39	0.39
Item (σ_{ζ_0})	0.58	0.58	0.55	0.56
Treatment (σ_{ζ_1})		0.05*	0.03	
Corr(Item,Treatment) (ρ_{12})		0.15	0.55	

Note: $^*p < 0.05$, $^{**}p < 0.01$.

Likelihood ratio tests revealed that Model 2 was a better fit than Model 1, and Model 3 was a better fit than Model 2, and Model 4 was not distinguishable from Model 3.

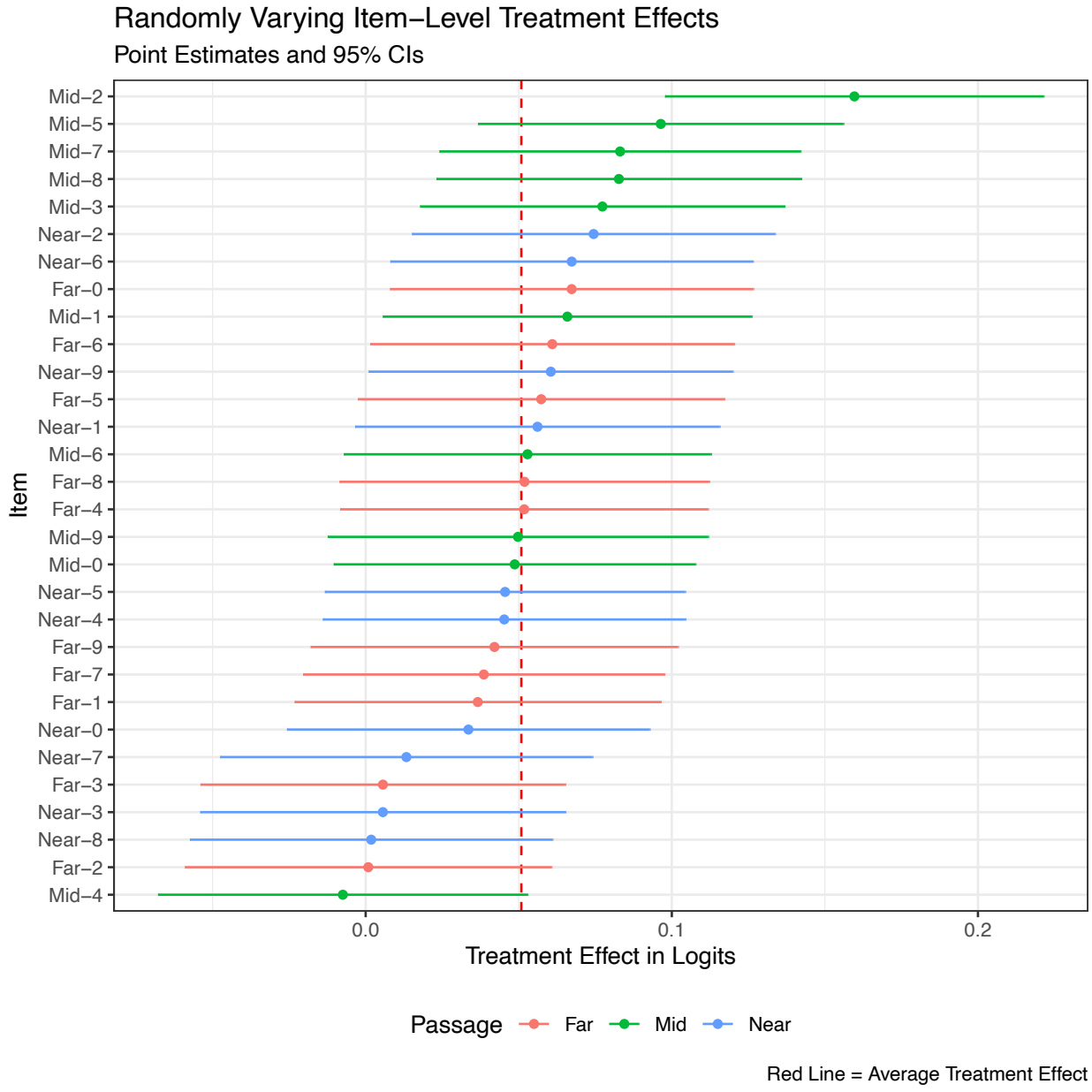
Likelihood ratio tests revealed that the item easiness-treatment effect size correlations in Models 2 and 3 were not statistically significant.

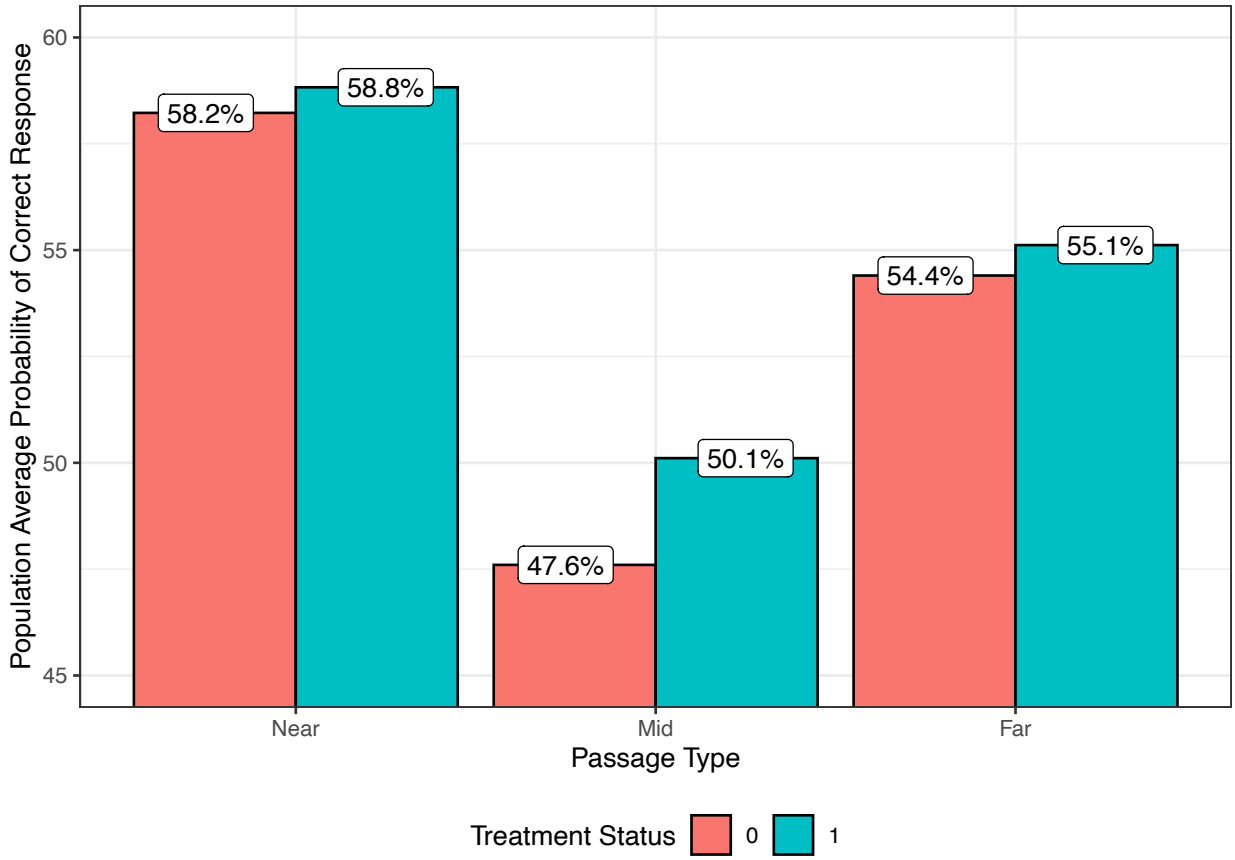
Pretest reading scores and main effects for each subtest were included in the model but omitted from the table.

Figure 3. Model-Implied Item- and Subtest-Level Treatment Effects

Top: Randomly Varying Item-Level Treatment Effects Color Coded by Subtest Passage Derived from Model 2

Bottom: Population Average Probabilities of Correct Response by Subtest Passage Type and Treatment Status Derived from Model 4





Discussion

Solely examining the average effect of an educational intervention may provide an incomplete picture of the efficacy of that intervention. A traditional statistical approach to examining HTE such as moderation or quantile regression attempts to explain variation in treatment effects as a function of person-level characteristics, as in moderation analysis, or the location of a subject in the conditional outcome distribution, as in quantile regression. While such methods are widely used and highly valuable, they ignore the potential HTE that may exist within an outcome measure itself. In contrast, the EIRM provides the ability to explore HTE from a new perspective, namely, the item level. Because the EIRM models all individual item responses directly, researchers can empirically estimate how much IL-HTE exists in the data by specifying a randomly varying item-level treatment effect in the model. Researchers can subsequently explore

to what extent treatment by item-characteristic interactions systematically explain the IL-HTE, and conversely, to what extent IL-HTE remains unexplained. Furthermore, estimates of the correlation between item easiness and treatment effect size may be of substantive interest to practitioners and applied researchers in understanding how an intervention affects student learning outcomes.

The results of this study clearly reveal several practical benefits to using the EIRM to model IL-HTE in practice. First, the simulation results show that even when IL-HTE are not present, allowing for them in the model does not materially affect the point estimates or standard errors of the average treatment effect. Second, when IL-HTE are present but not allowed for in the model, the standard errors associated with the average treatment effect are too small, providing an overly optimistic estimate of precision resulting in increased false positive rates with respect to the population average ATE. Therefore, researchers should test for IL-HTE when employing the EIRM to estimate treatment effects because it provides reasonably well-calibrated standard errors and false positive rates regardless of the true degree of item-level HTE, unless the researcher is interested only in the average treatment effect in the specific set of items on an assessment, in which case the constant effect model (or an item fixed effects model) is the appropriate substantive choice. Last, the application to the empirical reading comprehension assessment data from the MORE intervention showed that a null average treatment effect masked statistically significant and substantively meaningful IL-HTE. That is, rather than an ineffective intervention with a null effect, the EIRM revealed that MORE is most effective for the items of the Mid Transfer reading passage, a precise finding that may have been overlooked using other methods, suggesting that researchers should consider the possibility that interventions may differentially affect portions of a given outcome measure.

Limitations and Future Directions

While the potential value of examining IL-HTE through the EIRM is clear, the encouraging results of this study may be tempered by its simplifying assumptions. For example, the EIRM is typically estimated under the constraints of the One-Parameter Logistic (1PL) or Rasch model, in which all items are equally correlated with the latent trait. While the data generating process of this simulation was based on a 1PL model, a 1PL approach may not be appropriate for educational assessments in which items vary in their discriminations as well as their difficulties. Advances in estimation methods such as profile-likelihood (Jeon & Rabe-Hesketh, 2012) have enabled exploration of the Two-Parameter Logistic (2PL) EIRM that models item discriminations as either fixed quantities to be estimated, as in the `mirt` (Chalmers, 2012) or `PLmixed` (Rockwood & Jeon, 2018) R packages and the `gllamm` Stata program (Skrondal & Rabe-Hesketh, 2004), or as random variables to themselves be explained by the predictors in both frequentist (Petscher, et al., 2020, using Mplus; Cho, et al., 2014) and Bayesian paradigms (Bürkner, 2019, using R's `brms`). Similarly, the same unidimensionality and local independence assumptions of traditional IRT analysis also apply to the EIRM, and as such either the preliminary use of exploratory factor analysis before EIRM analysis (Petscher, et al., 2020, pp. 15-16) or the use of the multidimensional EIRM (De Boeck & Wilson, 2014) is recommended. Finally, application of the EIRM to non-dichotomous item responses would extend the utility of the EIRM to more diverse assessment contexts (Stanke & Bulut, 2019; Bulut, Gorgun, & Yildirim-Erbasli, 2021).

A final challenge of the application of the EIRM involves the interpretation of the coefficients of the fitted models. In contrast to a more familiar sum score, mean score, or standardized effect size, all but the most statistically literate practitioners are unlikely to have well-developed intuitions for the substantive meaning of treatment effect coefficients on the logit scale or the interpretational subtleties of logistic regression more generally (Mood, 2010), issues that

are compounded in the EIRM context by the difference between population-averaged (marginal) and cluster-specific (conditional) effects introduced by the cross-classified person- and item-level random effects of the parameterization (Austin & Merlo, 2017). As such, we suggest the following two approaches to make the EIRM results more interpretable. First, the fitted models can be used to estimate population-averaged response probabilities (e.g., using the `ggeffects` R package described in Lüdecke, 2018), as depicted earlier in the bottom panel of Figure 3, representing overall treatment-control contrasts on the probability scale that are likely to be more interpretable to stakeholders such as parents, teachers, or school leaders. Second, analysts can convert the EIRM treatment effect coefficient to a Cohen's *d* type effect size by the process of "y-standardization" (see Breen, Karlson, & Holm, 2018 for the single-level case; see Hox, Moerbeek, & Van de Schoot, 2017, Chapter 6 for the multilevel case), whereby the logit-scale coefficient β_{logit} is divided by the estimated total standard deviation of a postulated continuous variable Y* that could give rise to the observed dichotomous response Y, using the following formula

$$\beta_{ystd} = \frac{\beta_{logit}}{SD(Y^*)} = \frac{\beta_{logit}}{\sqrt{\frac{\pi^2}{3} + \sigma_\theta^2 + \sigma_{\zeta_0}^2 + \sigma_F^2}}$$

in which $\frac{\pi^2}{3} = 3.29$ is the variance of the logistic distribution, the σ_θ^2 and $\sigma_{\zeta_0}^2$ represent the variance components of the persons and items, and σ_F^2 is the variance of the fixed effects (i.e., the variance of the estimated linear predictor on the logit scale).

For the IL-HTE EIRM, the random slope associated with the treatment effect implies heteroskedasticity between the treatment and control groups (see Steele, 2008, pp. 29-32), with variances of

$$var(Y^*|treat = 0) = \frac{\pi^2}{3} + \sigma_\theta^2 + \sigma_{\zeta_0}^2 + \sigma_F^2$$

$$var(Y^*|treat = 1) = var(Y^*|treat = 0) + \sigma^2_{\zeta_1} + 2\sigma_{01}$$

Given the unequal variances when IL-HTE is present, we encourage standardizing by the control group to obtain a Glass's δ type effect size because IL-HTE will increase the variance of the treatment group, and therefore effects of otherwise equal magnitude would appear smaller due to the increased pooled SD as IL-HTE increases. The estimates from each group could be pooled if a Cohen's *d* type effect size were strongly desired.

While adding a layer of procedural complexity for the analyst, y-standardization has the advantage of (a) rendering logit coefficients comparable to those derived from linear regression with standardized continuous outcomes, (b) enabling comparison of multiple models fit to the same data and cross-sample comparisons of effect size (Breen, Karlson, & Holm, 2018), and (c) enabling the use of the effect size estimates in meta-analysis, contexts in which scale-free generalizability of the estimates is essential.

Conclusion

A principal aim of applied intervention research is to understand how far intervention effects travel. In this study, we leveraged online assessment data from a large-scale RCT to show how the impact of an evidence-based literacy intervention can promote transfer on a formative assessment of reading comprehension. In doing so, we simultaneously highlight the unique affordances of online assessments and the EIRM in identifying on what assessment tasks intervention effects emerge, thus illustrating how large-scale digital formative assessments can be leveraged to assess learning outcomes at scale across whole school systems. In sum, applying the EIRM to model IL-HTE can reveal a type of treatment impact variation to which other methods are blind. Data analysts can use the EIRM with varying item-level treatment effects to provide more insight for applied researchers by allowing more nuanced inference about the effects of

educational interventions on measured outcomes. In turn, more fine-grained findings will allow researchers to make more substantive and policy-relevant claims about intervention impacts, an approach that brings scholars one step closer to understanding for whom, under what conditions, and, crucially, on what assessment tasks an educational intervention works.

References

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

Bates D, Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01.

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: making an informed choice. *Quality & Quantity*, *53*(2), 1051-1074.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817-842.

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.

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. (2019). Bayesian item response modeling in R with brms and Stan. *arXiv preprint arXiv:1905.09501*.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, *48*, 1-29.

Chan, W., & Hedges, L. V. (2022). Pooling Interactions Into Error Terms in Multisite Experiments. *Journal of Educational and Behavioral Statistics*, 10769986221104800.

Cho, S. J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: Item modeling for explanation and item generation. *Psychometrika*, *79*(1), 84-104.

Christensen, K. B. (2006). From Rasch scores to regression. *Journal of Applied Measurement*, *7*(2), 184.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533-559.

De Boeck, P., & Wilson, M. (2014). 12 Multidimensional Explanatory Item Response Modeling. *Handbook of item response theory modeling: Applications to typical performance assessment*, 252.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach* (Vol. 10, pp. 978-1). New York: Springer.

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.

De Boeck, P., Cho, S. J., & Wilson, M. (2016). Explanatory item response models. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 249-266.

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

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

Jeon, M., & Rabe-Hesketh, S. (2012). Profile-likelihood approach for estimating generalized linear mixed models with factor structures. *Journal of Educational and Behavioral Statistics*, *37*(4), 518-542.

Jeon, M., & Rockwood, N. (2018). PLmixed: an R package for generalized linear mixed models with factor structures. *Applied Psychological Measurement*, *42*(5), 401.

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology*, *113*(1), 3.

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*. Advance online publication. https://psycnet.apa.org/doi/10.1037/edu0000751

Kim, Y. S., Petscher, Y., Foorman, B. R., & Zhou, C. (2010). The contributions of phonological awareness and letter-name knowledge to letter-sound acquisition—a cross-classified multilevel model approach. *Journal of educational psychology*, *102*(2), 313.

Lüdecke D (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software, 3*(26)*, 772*. doi: 10.21105/joss.00772

Miratrix, L. W., Weiss, M. J., & Henderson, B. (2021). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, *14*(1), 270-308.

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.

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*(2), 160-179.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475-499.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

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

Steele, F. (2008). Module 5: introduction to multilevel modelling concepts. *Centre for Multilevel Modelling*. Retrieved from: https://www.cmm.bris.ac.uk/lemma/pluginfile.php/306/mod_resource/content/2/C5%20Introduction%20to%20Multilevel%20Modelling.pdf

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. *Assessment of competencies in educational contexts*, 91-120.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*(4), 589-600.