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

Citation

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37378515

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Modeling Item-Level Heterogeneous Treatment Effects with the Explanatory Item Response Model: Leveraging Large Scale Online 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 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 Large Scale Online 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 systems (i.e., how systems function properly) was introduced 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 worked together 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 reading comprehension assessment.

The online assessment included three reading comprehension passages 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

\[
\logit \left( P(y_{ij} = 1) \right) = \beta_0 + \theta_j + \zeta_i
\]

\[
\theta_j \sim N(0, \sigma^2_{\theta})
\]

\[
\zeta_i \sim N(0, \sigma^2_{\zeta})
\]

in which the log-odds of a correct response to item \(i\) for person \(j\) is a function of the average log odds of a correct response (\(\beta_0\)), person ability (\(\theta_j\)) and item easiness (\(\zeta_i\)) (item easiness is the negative of what is often called item difficulty or location 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 analysts may choose 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 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 difficult than items that assess vocabulary), or both (e.g., do male-female performance gaps depend on item type).

The EIRM can also be used to model differential item functioning (DIF) (De Boeck, et al., 2011, pp. 18-19; Randall, Cheong, & Engelhard, 2011), that is, the phenomenon of respondents at the same level of the latent trait demonstrating different response probabilities for a specific item or cluster of items (AERA, 2014). Prior EIRM-based DIF analyses have demonstrated gender-based DIF in math assessments (Kan & Bulut, 2014), DIF for students with disabilities (Randall, Cheong, & Engelhard, 2011), or “instructional sensitivity” in longitudinal contexts (Naumann, Hochweber, & Hartig, 2014), among others. In the context of this study, IL-HTE can be conceptualized as uniform DIF, in that each residual item treatment effect represents differential performance between the treated and control groups above and beyond individual student ability ($\theta_j$) after any treatment effect that causes an overall difference in ability has been taken into
account. Therefore, we argue that the IL-HTE model we propose in this study conceptualizes DIF
as a feature of substantive interest reflecting the sensitivity of individual items to a treatment, rather
than a nuisance to be corrected for, for example, by removing items that demonstrate DIF from an
analysis. In sum, while the EIRM has primarily been applied to observational studies to examine
person- or item-level predictors of response patterns or DIF analyses, it can easily be extended 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:**

\[
\logit\left( P(y_{ij} = 1) \right) = \beta_0 + \beta_1 \text{treat}_j + \theta_j + \zeta_{0i}
\]

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

\[
\zeta_{0i} \sim N(0, \sigma_\zeta^2)
\]

**Model 2 – IL-HTE:**

\[
\logit\left( P(y_{ij} = 1) \right) = \beta_0 + \beta_1 \text{treat}_j + \theta_j + \zeta_{0i} + \zeta_{1i} \text{treat}_j
\]

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

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

in which \(y_{ij}\) is the dichotomous response to item \(i\) for person \(j\), \(\beta_0\) is the log-odds of a correct
response for a student of average ability to an item of average difficulty, \(\beta_1\) is the average treatment
effect (ATE) across items, \(\theta_j\) is a random intercept representing unexplained person ability (as in
traditional IRT modeling), and $\xi_{0i}$ 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, $\xi_{1i}$, that captures the deviation between each item’s individual treatment effect and the ATE $\beta_1$, thus allowing for IL-HTE. Model 2 can also allow for correlation between item easiness and item treatment effect size ($\rho_{01}$) if it is hypothesized that the treatment would have larger impacts on, for example, the easiest items on the assessment. Such a model could then be tested against a model in which the correlation is constrained to equal zero to determine if the correlation is necessary. 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. Note that this approach to IL-HTE analysis can be performed simultaneously with other HTE methods, for example, by including person characteristic by treatment interactions in the model, but for clarity of exposition, we explore only IL-HTE for the purposes of this study.

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 effects 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 easiness 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 $2 \times 3$ 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.

The specific focus of this study on IL-HTE thus builds on existing simulation literature on the EIRM and associated models, both generally (Christensen, 2006; Zwinderman, 1991; Christensen, et al., 2004), and in longitudinal (Wilson, Zheng, & McGuire, 2012), instrumental variables (Rabbitt, 2018), and Bayesian (Lozano & Revuelta, 2021) settings, all of which have generally found that item parameters and regression coefficients are well recovered by the EIRM when sample sizes are large.

**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 representing varying degrees of transfer from the MORE curriculum, and all students received the same set of items. The assessment was administered online at the end of the MORE intervention, but prior to the high-stakes end-of-year state test.

In this study, learning transfer was conceptualized along a continuum from near transfer to far transfer (Barnett & Ceci, 2002). Accordingly, we developed a transfer assessment of science content reading comprehension to measure third graders’ ability to comprehend texts about topics related to the schema for how living systems function properly and how leaders in social systems support scientific innovations. In particular, the MORE science lessons delivered to all students focused on the specific topic of the skeletal, muscular, and nervous systems of the human body. The social studies extension lessons provided only to treatment students discussed the development of the Apollo 11 Moon Mission. Transfer was defined by the number of explicitly taught domain specific vocabulary words contained in each passage, in that the Near-transfer passage contained seven explicitly taught vocabulary words, Mid contained four, and Far contained none. The three reading passage topics included scientists studying monkey hearts (Near transfer; seven taught vocabulary words), the exploration of Mars (Mid transfer; four taught vocabulary words), and the search for the Titanic (Far transfer; no taught vocabulary words). Thus,
we predicted that the Near transfer passage would be relatively easier than the Mid and Far Transfer passages for all students due to the presence of untaught vocabulary words.

The primary substantive research aim of this analysis 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 general schema for system developed through the social studies extension lessons while reading the Mid and Far Transfer passages, we hypothesized that treatment students would potentially outperform control students on Mid and Far Transfer items. While findings from a previous RCT involving a similarly structured assessment with second grade students demonstrated the largest treatment effects on the Near transfer items among students who were assigned to receive MORE lessons compared to business-as-usual controls (Kim, et al., 2022), we hypothesized that the intervention impacts of this study be instead most pronounced on the Mid and Far transfer passages, because these passages contained social studies concepts that only treatment students were exposed to. Furthermore, because the topic of the Mid transfer passage (i.e., the exploration of Mars) was closely conceptually related to the content of the social studies extension read aloud lessons provided to the treatment students (i.e., the Apollo 11 Moon Mission), we predicted the largest treatment effects on the Mid transfer passage.

The psychometric properties of the assessment, including IRT item characteristic curve plots, exploratory factor analysis (EFA) scree plots, confirmatory factor analysis (CFA) fit statistics, and additional descriptive statistics are presented in the Online Supplemental Materials
one item was removed due to poor functioning (i.e., a negative discrimination parameter) in a 2PL IRT analysis, resulting in 29 items retained for the analyses presented here. Internal consistency was estimated at 0.80, and EFA provided strong evidence of unidimensionality, with the first factor explaining far more total variance than subsequent factors, and fit statistics for the unidimensional CFA model were strong (CFI = 0.94, TLI = 0.94, RMSEA = 0.031, SRMR = 0.025), suggesting that the application of the unidimensional EIRM is justifiable.

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:**

\[
\text{logit}(P(y_{ijk} = 1)) = \beta_0 + \beta_1 \text{treat}_k + \beta_2 \text{pretest}_{jk} + \theta_{jk} + \zeta_{0i} + \nu_k
\]

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

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

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

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

\[
\text{logit}(P(y_{ijk} = 1)) = \beta_0 + \beta_1 \text{treat}_k + \beta_2 \text{pretest}_{jk} + \theta_{jk} + \zeta_{0i} + \zeta_{1i} \text{treat}_k + \nu_k
\]

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

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

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

**Model 3 – MORE Assessment EIRM 3, Systematically and Randomly Varying IL-HTE:**
\[
\logit\left( P(y_{ijk} = 1) \right) = \beta_0 + \beta_1 \text{treat}_k + \beta_2 \text{pretest}_{jk} + \beta_3 \text{mid}_i + \beta_4 \text{far}_i + \beta_5 \text{treat}_k \times \text{mid}_i \\
+ \beta_6 \text{treat} \times \text{far}_i + \theta_{jk} + \zeta_{0i} + \zeta_{11} \text{treat}_k + \nu_k
\]

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

\[
\nu_k \sim N\left(0, \sigma_{\nu}^2\right)
\]

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

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

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

\[
\nu_k \sim N\left(0, \sigma_{\nu}^2\right)
\]

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 \((\nu_k)\) to account for the cluster-randomized design, a student-level reading pretest score\(^1\) to improve the precision of the estimates \((\beta_2)\), main effects for passage easiness \((\beta_3, \beta_4)\), and passage by treatment interactions \((\beta_5, \beta_6)\) to model systematic sources of IL-HTE. Across all EIRMs, we

---

\(^1\) The pretest employed in this study is the NWEA Measure of Academic Progress (MAP) reading assessment, a state-mandated test administered at the beginning of the school year. While the inclusion of the pretest variable makes the MORE impact models less analogous to the simpler models evaluated in the simulations, the gains to precision through inclusion of the pretest were substantial in that they resulted in a 40% reduction in the standard error of the treatment coefficient.
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. The systematic model-building strategy outlined above provides a template for exploring the presence, magnitude, and predictors of IL-HTE, but researchers with strong *a priori* hypotheses may instead elect to fit a single model that includes, for example, randomly and systematically varying treatment effects at the item level.

**Results**

**Monte Carlo Simulation**

The results of the simulation reveal first that the point estimates for average treatment effects for the constant (i.e., finite sample) and IL-HTE (i.e., population) models are nearly identical ($r = 0.999$), and that the bias associated with the treatment effect parameter $\beta_1$ was negligible across all conditions, at less than 0.01 logits. Analysis of the uncertainty associated with the point estimates is more revealing of the differences between the models. 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 across simulation trials$^2$) 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 IL-HTE model (i.e., within ±5% of the true value, as indicated by the gray bands in the figure), but the model-based SEs for the constant treatment effect model are systematically too low when IL-HTE is high, falling more than 5% below the diagonal identity

---

$^2$ The population average SE is calculated as the standard deviation of the point estimates around the mean $\beta_1$, in which $n$ indexes the number of simulations: $\sqrt{\frac{\sum(\hat{\beta}_1 - \beta_1)^2}{n-1}}$
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 point estimates with respect to 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 both a statistical issue that can be investigated empirically, and a substantive concern regarding whether to view the test being used as a representative sample of an item bank describing the latent construct of interest, 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

*Bottom:* Test-specific (i.e., finite sample) estimand

Dotted Lines show ±5% from True Value
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 indistinguishable from 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). Finally, though the simulation models are simpler than the empirical models, supplementary analyses in the OSM indicate that the pattern of findings is essentially identical when the data generating process is based on a cluster randomized trial that includes a student-level pretest covariate, an item-level predictor, and HTE at the cluster level (though standard error calibration appears worse for all models in the cluster randomized context), suggesting that the general pattern of results presented here is likely to generalize to other data analytic settings.
Figure 2: Comparison of false positive rates by method based on true IL-HTE

Note. Confidence intervals for the false positive rates were calculated using the standard formula for the standard error of a proportion, $\sqrt{\frac{pq}{n}}$, where $n$ indexes the number of simulation trials.

**Application to Empirical MORE 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 (assuming that item-level treatment effects are constant, i.e., the finite sample ATE) is positive but not statistically significant ($\beta_1 = 0.06, SE = 0.09$). 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 around the population average treatment effect, as the SD of the randomly varying item-level treatment effect ($\sigma_{\xi_1} = 0.048, p < 0.05$) is as large as the point estimate itself ($\beta_1 = 0.06$) and implies a 95% prediction interval of approximately -0.04 to +0.16 for individual item-level treatment effects on the logit scale.
Importantly, even without treatment by item type interactions to explain the IL-HTE variance, Model 2 may still be preferable to Model 1 given the simulation results, because it targets the treatment effect in the population of items, provides better calibrated SEs (and therefore p-values and false positive rates), allows for the generation of a 95% prediction interval of individual item treatment effects, and can serve as an exploratory procedure to generate hypotheses about potential item-level predictors of treatment effect size.

Model 3 tests the hypothesis that item easiness and treatment effects systematically depend on the passage type by including both passage main effects and treatment by passage type interaction terms. The main effects of passage type in Model 3 confirm that the Mid and Far transfer passages were slightly more difficult than the Near transfer passage (for control students), as expected, but these differences are not statistically significant. 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 ($\beta_5 = 0.10$, $p < 0.001$). 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 both science and social studies vocabulary words and was most closely thematically related to the treatment group intervention lessons.

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 when comparing Model 2 to Model 3, the main effects for passage type ($\beta_3, \beta_4$) and the treatment by passage type interaction terms ($\beta_5, \beta_6$) explain approximately 7% of the item easiness variance (.3428 to .3179) and approximately 87% of the treatment effect variance (.0023 to .0003).
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, as an analysis that used the estimated theta score from a 2PL IRT model as a continuous outcome revealed a standardized effect size of 0.05 and was non-significant (SE = 0.06, z = 0.83).

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

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Mod. 1</th>
<th>Mod. 2</th>
<th>Mod. 3</th>
<th>Mod. 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\beta_0$)</td>
<td>0.16</td>
<td>0.16</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>1 = Treatment ($\beta_1$)</td>
<td>0.06</td>
<td>0.06</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Pretest (Std.) ($\beta_2$)</td>
<td>0.91 ***</td>
<td>0.91 ***</td>
<td>0.91 ***</td>
<td>0.91 ***</td>
</tr>
<tr>
<td>1 = Mid Passage ($\beta_3$)</td>
<td></td>
<td></td>
<td>-0.38</td>
<td>-0.38</td>
</tr>
<tr>
<td>1 = Far Passage ($\beta_4$)</td>
<td></td>
<td></td>
<td>-0.16</td>
<td>-0.16</td>
</tr>
<tr>
<td>1 = Treatment x Mid ($\beta_5$)</td>
<td>0.10 ***</td>
<td>0.10 ***</td>
<td>0.10 ***</td>
<td>0.10 ***</td>
</tr>
<tr>
<td>1 = Treatment x Far ($\beta_6$)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Random Effects

<table>
<thead>
<tr>
<th></th>
<th>Mod. 1</th>
<th>Mod. 2</th>
<th>Mod. 3</th>
<th>Mod. 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student ($\sigma_0^2$)</td>
<td>0.4527</td>
<td>0.4528</td>
<td>0.4528</td>
<td>0.4528</td>
</tr>
<tr>
<td>School ($\sigma_s^2$)</td>
<td>0.1895</td>
<td>0.1896</td>
<td>0.1896</td>
<td>0.1898</td>
</tr>
<tr>
<td>Item ($\sigma_i^2$)</td>
<td>0.3435</td>
<td>0.3428</td>
<td>0.3179</td>
<td>0.3246</td>
</tr>
<tr>
<td>Treatment ($\sigma_t^2$)</td>
<td>0.0023</td>
<td>0.0003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cov(Item, Treat)</td>
<td>0.0055</td>
<td>0.6676</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC</td>
<td>0.2305</td>
<td>0.2307</td>
<td>0.2272</td>
<td>0.2272</td>
</tr>
<tr>
<td>N Students</td>
<td>7797</td>
<td>7797</td>
<td>7797</td>
<td>7797</td>
</tr>
<tr>
<td>N Items</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>N Schools</td>
<td>110</td>
<td>110</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td>Observations</td>
<td>219252</td>
<td>219252</td>
<td>219252</td>
<td>219252</td>
</tr>
<tr>
<td>Deviance</td>
<td>243828.716</td>
<td>243822.673</td>
<td>243806.095</td>
<td>243807.393</td>
</tr>
</tbody>
</table>

Note. Likelihood ratio tests revealed that Model 2 was a better fit than Model 1, Model 3 was better than Model 2, and Model 4 was indistinguishable from Model 3. LR tests revealed that the item easiness-treatment effect size correlations in Models 2 and 3 were not statistically significant.
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
Note 1. The confidence intervals of the random slope residuals were calculated using the normal approximation of ±1.96 times the posterior standard error. This figure should be taken as somewhat approximate, as due to the shrinkage of the Empirical Bayes estimate, coverage is not guaranteed. Under a fully Bayesian model these could be interpreted as posterior intervals; given a large sample size, this approximation will generally be good.

Note 2. When the potentially outlying item Mid-2 is removed from the analysis as a sensitivity check, we find that the treatment by mid transfer passage interaction effect is slightly smaller in magnitude but still statistically significant in Models 3 and 4, but the random slope for treatment is no longer significant in Model 2.

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, as (a) the correlation between the point estimates for the two methods was near perfect \( r = 0.999 \), (b) the bias associated with the treatment effect parameter \( \beta_1 \) was negligible across all conditions, at less than 0.01 logits, and (c) the standard errors for the IL-HTE model are well-calibrated even when the true level of IL-HTE is 0, as displayed in Figure 1. 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 (i.e., with a likelihood ratio test comparing models with and without IL-HTE) when employing the EIRM to estimate treatment effects because it provides 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, random intercepts model (or an item fixed effects model, see Miratrix, et al., 2021, p. 280) 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). However, sensitivity analyses based on a 2PL data generating process presented in the OSM demonstrate that varying item discriminations have no impact on treatment effect bias, false positive rates, or statistical power, but do result in slightly less accurate standard errors (though still within approximately 5% of their true values). Therefore, the 1PL EIRM appears robust to this type of misspecification, which is an important finding given that the 2PL EIRM has been demonstrated to suffer from accuracy issues (Zhang, Ackerman, & Wang 2021). 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. Furthermore, 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). We also find that the EIRM requires relatively large sample sizes and highly variable treatment effects for meaningful IL-HTE analysis, as additional simulations to determine statistical power presented in the OSM indicate that 80% power for detecting IL-HTE under individual randomization is only achieved with at least 300 subjects, 20 items, and a large treatment effect standard deviation of 0.40 logits, and as such is best suited for analysis of relatively large data sets.

An additional 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 may be more interpretable to stakeholders such as parents, teachers, or school leaders, showing a small but statistically significant treatment effect of 3.1 percentage points on the Mid Transfer passage items. Similar

---

3 In the cluster randomized MORE data, the number of students is 7797, number of schools is 110, the average number of students per school is 70.8, and the school-level intraclass correlation (ICC) from an unconditional EIRM is 0.108. Applying an effective sample size (ESS) calculation (making the simplifying assumption of an equal number of subjects per cluster) yields an ESS of \( \frac{7797}{1 + 0.108(70.8 - 1)} \approx 913 \) units (Killip, Mahfoud, & Pearce, 2004), suggesting a sufficient sample size for IL-HTE analysis, particularly considering the inclusion of a strong pretest covariate in the models and the assessment length of 29 items.
alternative metrics to facilitate communication to non-technical audiences include relative percentile ranks, conditional probabilities of a correct response compared to a specified baseline (e.g., 50%), or odds ratios, depending on the intended audience. 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 $\beta_{\logit}$ 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 $\sigma_\theta^2$ and $\sigma_{\zeta_0}^2$ represent the variance components of the persons and items, and $\sigma_F^2$ is the variance of the fixed effects (i.e., the variance of the estimated linear predictor on the logit scale)$^4$.

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

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

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

Given the unequal variances when IL-HTE is present, we encourage standardizing by the control group to obtain a Glass’s $\delta$ type effect size because IL-HTE will increase the variance of the

---

$^4$ When there are no covariates in the model beyond the treatment group indicator, and the model has no level of nesting above persons, the estimated standard deviation of the person random effects (i.e., $\sigma_\theta$) could be used as the denominator in the calculation of a Cohen’s $d$ type effect size, as it represents the pooled within-group standard deviation of $\theta$. 
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. For example, applying the above formula to the MORE assessment data yields a latent control group standard deviation of $2.45 (\sqrt{3.29 + 0.45 + 0.19 + 0.32 + 1.76})$, resulting in a treatment effect size equivalent of approximately $0.053$ standard deviations for the Mid Transfer passage $\left(\frac{0.03+0.10}{2.45}\right)$, a small but meaningful positive effect given the brevity of the intervention.

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 an assessment of reading comprehension. In doing so, we simultaneously highlight the affordances of online assessments (e.g., low cost and high accuracy at large scale) and the EIRM in identifying on what assessment tasks intervention effects emerge, thus illustrating how large-scale digital 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


engagement through a first-grade content literacy intervention. *Journal of Educational Psychology, 113*(1), 3.


