




Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in R

Joshua B. Gilbert¹ 

Accepted: 16 September 2023
© The Psychonomic Society, Inc. 2023

Abstract

Recent advancements in education scholarship have introduced Item Response Theory (IRT) models to address treatment heterogeneity at the assessment item level. These models for item-level heterogeneous treatment effects (IL-HTE) enable detailed analyses of treatments that may have varying impacts on individual items within an assessment. This article offers a comprehensive tutorial for applied researchers interested in implementing IL-HTE analysis in R, utilizing the lme4 package. Using empirical data from a second-grade reading comprehension assessment as a running example, this tutorial emphasizes model-building strategies, interpretation techniques, visualization methods, and extensions. By following this tutorial, researchers will gain practical insights into utilizing IL-HTE analysis for enhanced understanding and interpretation of treatment effects at the item level.

Keywords Causal inference · Treatment heterogeneity · Explanatory item response model · Educational measurement · R · lme4

Introduction

Researchers in the social sciences are increasingly drawn to models for treatment heterogeneity due to their ability to provide policy-relevant findings by capturing systematic variation in treatment effects. In recent education scholarship, a novel approach has emerged for exploring treatment heterogeneity by leveraging item response theory (IRT) models. These models offer a means to evaluate treatment effects that differ at the assessment item level, where specific assessment items may respond distinctively to the treatment. By employing IRT models for treatment heterogeneity, researchers can gain deeper insights into treatment effects and their impact on individual assessment items, item clusters, or test subscales (Ahmed et al., 2023; Gilbert, Kim, & Miratrix, 2023).

These “item-level heterogeneous treatment effects” (IL-HTE) (Gilbert et al., 2023) “item-treatment interactions” (Ahmed et al., 2023), or “item-specific effects” (Sales et al., 2021) are both statistically and practically significant. Prior

research employing IL-HTE models is rare, but has nonetheless shown promising applications. Gilbert et al. (2023) applied the IL-HTE model to assess transfer effects on a reading comprehension assessment of third grade students and found significant positive treatment effects on the item cluster associated with one reading passage and null effects on two others, a finding that would have been masked in a traditional analysis. Sales et al. (2021) used the IL-HTE model to determine whether assessment items were differentially impacted by tutoring and found that items most closely related to the tutoring content did not show larger effects, which has potential implications for the analysis of test score inflation (Koretz, 2005) and the alignment between interventions and assessments (Francis, Kulesz, Khalaf, Walczak & Vaughn, 2022). Ahmed et al. (2023) examined item-level data from 15 randomized controlled trials (RCTs) containing over seven million item responses and found evidence that IL-HTE is common in applied interventions, including a case where the overall treatment effect was mostly driven by a single test item. From a statistical perspective, Gilbert et al. (2023) demonstrated that failing to account for IL-HTE when it is present in the data results in underestimated standard errors and invalid hypothesis tests and showed that null average effects can mask positive item-specific or subscale treatment effects. Sales et al. (2021) showed that the Empir-

✉ Joshua B. Gilbert
joshua_gilbert@g.harvard.edu

¹ Harvard Graduate School of Education, 13 Appian Way,
Cambridge 02138, MA, USA

ical Bayes estimates of item-specific treatment effects are robust to multiple comparisons and thus provide a useful tool to identify individual items that are particularly sensitive to treatment.

While this newly emerging literature clearly articulates the statistical and substantive rationale for IL-HTE models, there are no tutorials for applied researchers in conducting IL-HTE analysis. The purpose of this paper is to provide such a tutorial for researchers in R using the `glmer` function from the `lme4` package, emphasizing model-building strategies, parameter interpretation, model visualization, and extensions with a worked example from education research. As such, we build on the work of De Boeck et al. (2011) and the earlier work of Doran, Bates, Bliese and Dowling (2007), who demonstrated how to use `lme4` to fit descriptive and explanatory IRT models and extend their approach to model IL-HTE.

Traditional models for HTE

To best understand the affordances of the IL-HTE model, let us first consider the most typical approach to HTE analysis, that is, treatment by subject characteristic interactions. Consider the following model:

$$E(Y_i) = \beta_0 + \beta_1 \text{treat}_i + \beta_2 X_i + \beta_3 \text{treat} \times X_i$$

which Y_i is some outcome variable for person i , treat_i is an indicator variable for treatment status, X_i is a covariate of interest (e.g., age, gender, prior ability, etc.) and $\text{treat} \times X_i$ is the interaction between the treatment and the covariate. When $\beta_3 = 0$, the average treatment effect (ATE) is constant—it is β_1 —across the range of X . When $\beta_3 \neq 0$, the magnitude of the ATE is conditional on the value of X . In particular, the conditional ATE (CATE) is equal to $\beta_1 + \beta_3 X_i$. In other words, under this simple model, the ATE varies as a function of X and the HTE is fully explained by person characteristic X .

However, researchers may be interested in other sources of HTE. For example, if Y_i represents an educational test score comprised of several components, it is conceivable that the treatment has differential impacts on each of these components. As long as Y_i is treated as a single numeric quantity in an analysis, however, such “within-outcome” HTE would not be detectable by the model above. For example, if Y_i represents a sum or factor score on a math test, we may be interested in whether treatment is more effective for geometry or algebra items. To fit such a model in the standard regression framework, the analyst would have to fit two separate regression models for each subscale. While such an approach has merit, it also is weakened by several limitations. First, it requires an *a priori* specification of which subscales the outcome measure should be separately analyzed. Second, it assumes that within each subscale, itself comprised

of several test items, the treatment effect is constant. Third, each subscale is less reliable than the overall score because it has fewer items and lower reliability translates to greater attenuation bias when outcome variables are standardized (Gilbert, 2023; Hedges, 1981). Finally, when separate outcome models are fit, there is no direct test of the *difference* in treatment effects between the two models, which is often of substantive interest. While many of these limitations have workarounds, such as errors-in-variables models or structural equation modeling approaches, we will show that the IL-HTE model provides a parsimonious approach to testing for item-specific or subscale treatment effects.

The IL-HTE model

The IL-HTE model is an extension of the Explanatory Item Response Model (EIRM) (Wilson & De Boeck, 2004), a flexible approach that combines IRT (i.e., measurement, psychometric, latent variable) models and regression models into a single estimation procedure. The EIRM emerged from earlier work that synthesized IRT models and mixed effects models into a unified framework (Adams, Wilson & Wang, 1997; Rijmen, Tuerlinckx, De Boeck & Kuppens, 2003). In its most basic form, the EIRM relates an item response probability to a nonlinear function of person and item parameters. Consider first the dichotomous case where item responses can be correct (1) or incorrect (0). We define η_{ij} as the log-odds of the probability that observed response Y to item i from person j is equal to 1:

$$\eta_{ij} = \text{logit}(P(Y_{ij} = 1)).$$

η_{ij} is in turn a linear function of person ability θ_j and item location (commonly interpreted as “easiness” on an educational test) b_i ¹:

$$\eta_{ij} = \theta_j + b_i$$

$$\theta_j = \beta_0 + \varepsilon_j$$

$$b_i = \zeta_{0i}$$

$$\varepsilon_j \sim N(0, \sigma_\varepsilon)$$

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

θ_j can be expressed as average ability β_0 and normal deviations from that average ε_j , and the item easiness parameters ζ_{0i} are normally distributed. The EIRM is therefore a special case of cross-classified logistic regression, in which the item responses are nested in the cross-classification of persons and items. The EIRM can also be considered a special case of the

¹ On an educational test, b_i is most often interpreted as item easiness. Item easiness is the negative of what is usually called item difficulty in the IRT literature. Because `lmer` syntax allows most easily for the estimation of item easiness parameters, and our examples are drawn from education research, we will proceed with the item easiness terminology.

generalized linear mixed model (GLMM). With no predictors in the model, the EIRM is equivalent to a one-parameter logistic (1PL) or Rasch IRT model, with random item easiness parameters (De Boeck, 2008). While item parameters are traditionally considered fixed in IRT analysis, the random effects specification offers several advantages in modeling IL-HTE because it (a) provides estimates of uncertainty that correspond to the (real or hypothetical) population of items from which a test was constructed, (b) allows for the inclusion of item predictors (which would be collinear with item indicators in an item fixed effects model), and (c) provides a direct parameter estimate of the amount of IL-HTE in the data (Gilbert et al., 2023).

Both person and item predictors can be added to the EIRM to explain systematic differences in person ability or in item easiness. For example, consider the causal inference case with a person-level treatment $treat_j$, added to the equation for θ_j :

$$\begin{aligned}\eta_{ij} &= \theta_j + b_i \\ \theta_j &= \beta_0 + \beta_1 treat_j + \varepsilon_j \\ b_i &= \zeta_{0i} \\ \zeta_{0i} &\sim N(0, \sigma_{\zeta 0}) \\ \varepsilon_j &\sim N(0, \sigma_e)\end{aligned}$$

Here, β_1 represents the average treatment effect (ATE) on the latent trait θ_j and β_0 represents the mean of the latent trait in the control group. That is, treatment (hopefully) increases the latent ability for treated students, which would result in an increased logit of a correct response across all items. The ATE is implicitly assumed constant across the items. The EIRM with person predictors has a long history (Wilson & De Boeck, 2004), but applications have historically been limited to descriptive, rather than causal, analyses (Gilbert et al., 2023).

We can allow the treatment effect to vary across the items by adding a random slope term for treatment in the equation for b_{ij} (wherein the ij subscript allows for separate b parameters for each treatment group), resulting in the IL-HTE model:

$$\begin{aligned}\eta_{ij} &= \theta_j + b_i \\ \theta_j &= \beta_0 + \beta_1 treat_j + \varepsilon_j \\ b_{ij} &= \zeta_{0i} + \zeta_{1i} treat_j \\ \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} &\sim N\left(0, \begin{bmatrix} \sigma_{\zeta 0} & \rho \\ \rho & \sigma_{\zeta 1} \end{bmatrix}\right) \\ \varepsilon_j &\sim N(0, \sigma_e)\end{aligned}$$

As before, β_1 represents the ATE, but it now represents the treatment effect on the *average* item, and each item has a unique treatment effect that deviates from this average. That is, the treatment effect for item i can be decomposed into the ATE β_1 plus the item-specific residual ζ_{1i} , such that larger

values of ζ_{1i} imply larger item-specific treatment effects. The variance of these item-specific deviations from the ATE is parameterized by the standard deviation (or variance) of ζ_{1i} , $\sigma_{\zeta 1}$, thus providing a direct estimate of the amount of IL-HTE in the data. Item-specific treatment effects and item easiness parameters may be correlated, captured by the parameter ρ . A statistically significant value for $\sigma_{\zeta 1}$ indicates that the treatment effect is *not* constant across all items, but varies randomly about the ATE β_1 . In other words, the IL-HTE model allows the researcher to determine whether “observed treatment effects [are] due to generalized gains on the aggregate achievement measures or are ... due to targeted gains on specific items” (Ahmed et al., 2023).

The analyst may then add treatment by item-characteristic interaction terms to the model to explain some of the item-level treatment effect variance to capture systematic IL-HTE. For example, if a mathematics assessment contains algebra and geometry items, a treatment by item type interaction would allow for the treatment effect to differ by the type of item, even if the assessment is a unidimensional measure of mathematics proficiency. Understood in these terms, IL-HTE can equivalently be conceptualized as a type of uniform differential item functioning (DIF), in that each ζ_{1i} represents an additional treatment effect on item i above and beyond any overall treatment effect on θ_j (i.e., β_1) has been taken into account (Montoya & Jeon, 2019). Thus, IL-HTE may reflect the “instructional sensitivity” (Naumann, Hochweber & Hartig, 2014) of an item, rather than a defect of the assessment instrument prompting the removal of the items exhibiting DIF (Gilbert et al., 2023). In other words, the IL-HTE framework conceptualizes DIF as a feature of substantive interest rather than a nuisance to be corrected for. Even though IL-HTE implies a violation of measurement invariance, we argue that the ability of IL-HTE analysis to provide novel insights into empirical assessment data outweighs the violation of classical measurement principles it implies.

Empirical data

To provide a working example for our IL-HTE tutorial, we employ a public use file from Kim et al. (2023). The authors examined a sustained content literacy intervention called the Model of Reading Engagement (MORE) and estimated its effects on a researcher-designed reading comprehension assessment. Thirty schools were randomly assigned to implement the MORE intervention curriculum in first and second grade ($N = 2174$), and all students completed the assessment at the end of second grade. The assessment contained twenty dichotomously scored items across three reading passages (overall $\alpha = .78$), representing varying degrees of transfer from the MORE intervention curriculum. That is, the “near transfer” passage was designed to most closely track the vocabulary and concepts taught in MORE, and the “mid”

and “far transfer” passages were designed to represent concepts and vocabulary more distant from MORE. The authors tested the hypothesis that treatment effects differed according to the passage type, with the prediction that the effects would be largest on the “near transfer” passage, smaller on the “mid transfer” passage, and smallest on the “far transfer” passage. Here, we replicate and extend their analysis to illustrate the affordances of the IL-HTE model and its implementation in R. For clarity of exposition, we ignore the school-level clustering in the data and omit many of the demographic covariates the authors employed to maintain focus on the model building process and interpretation. We direct the reader to the original paper for more detail, but note that we find essentially identical results to those reported by Kim et al. (2023) under the models explored here.

The code below loads the data from Kim et al. (2023) into R using the `dataverse` package. First, we load the relevant libraries for the tutorial. Users may need to install the R packages using the `install.package("package_name")` syntax. The resulting data set contains the following six variables:

1. `s_id`: anonymous student identifier
2. `treat`: 0/1 treatment indicator
3. `item`: item number (1–20)
4. `s_correct`: 0/1 indicator for a correct response to the item
5. `pretest`: baseline (pre-intervention) state reading test scores, standardized
6. `passage`: factor variable indicating transfer passage type (near, mid, far)

```
# load libraries
library(tidyverse) # data manipulation
library(lme4) # multilevel models
library(dataverse) # dataverse access
library(broom.mixed) # tidying functions
library(ggeffects) # convenience
function for graphing
library(brms) # bayesian models
library(texreg) # regression tables
# set ggplot theme
theme_set(theme_classic())
# read in the public files and limit
to relevant variables
# set the dataverse server
Sys.setenv("DATAVERSE_SERVER" =
"dataverse.harvard.edu")
# bring in the student level data
more_long <- get_dataframe_by_name(
  filename = "item_response_public.tab",
  dataset = "https://doi.org/10.7910/DVN
/LAWFFU"
```

```
) |>
  select(s_id, treat = s_itt_consented,
item = s_q_num,
        s_correct, s_mid, s_far,
        pretest = s_maprit_1819w_std) |>
  # order by ID and question number
  arrange(s_id, item) |>
  # create a factor for the passage type
  mutate(passage =
case_when (
  s_mid == 1 ~ "mid",
  s_far == 1 ~ "far",
  .default = "near"
),
  passage = factor(passage, levels = c(
"near", "mid", "far"))) |>
  select(-s_mid, -s_far)
```

We examine the first six rows of the data set:

```
head(more_long)

## # A tibble: 6 x 6
##       s_id treat item s_correct pretest passage
##       <dbl> <dbl> <dbl>     <dbl>   <dbl>   <fct>
## 1 110027826921 0     1         1 -0.180 near
## 2 110027826921 0     2         0 -0.180 near
## 3 110027826921 0     3         0 -0.180 near
## 4 110027826921 0     4         0 -0.180 near
## 5 110027826921 0     5         0 -0.180 near
## 6 110027826921 0     6         1 -0.180 near
```

Fitting the IL-HTE model

We systematically build the IL-HTE model to illustrate the method and emphasize the statistical and substantive interpretation of the model results. We begin with a baseline model without predictors, then add a constant treatment effect, proceed with an examination the IL-HTE model with and without treatment by item-type interaction terms, and conclude with extensions that demonstrate the applicability of the IL-HTE model to more diverse contexts.

Model 0: Baseline

We begin with the unconditional Rasch model with random item easiness parameters as a baseline for comparison and to estimate the unconditional variances of person and item parameters in the data set:

$$\begin{aligned}\eta_{ij} &= \theta_j + b_i \\ \theta_j &= \beta_0 + \varepsilon_j \\ b_i &= \zeta_{0i} \\ \varepsilon_j &\sim N(0, \sigma_e) \\ \zeta_{0i} &\sim N(0, \sigma_{\zeta_0})\end{aligned}$$

To fit the model, we use the `glmer` function from `lme4`, which allows for the estimation of GLMMs, including the EIRM. The syntax is straightforward, in that we specify our outcome variable `s_correct` as a function of the intercept 1, a random effect for students (`1|s_id`), and a random effect for items (`1|item`). We declare the data using the argument `data = more_long` and the link function using `family = binomial` in which the logistic model is the default. The argument `nAGQ` specifies the number of integration points used in the adaptive Gauss-Hermite Quadrature of the estimation procedure. `nAGQ = 1` is the default for `glmer` models and corresponds to the Laplace approximation. Higher values for `nAGQ` will provide more accurate results but require greater computational time because of the numerical integration required in estimation, particularly when there are many random effects. Because computation time increases exponentially with each additional integration point, Rabe-Hesketh and Skrondal (2022) suggest that analysts start with a small number of integration points and test sequentially higher numbers on a simple model; if the results are substantially similar, the lower number is likely to be adequate as the model building proceeds (see pp. 596–602 for a full discussion). For the purposes of this tutorial, we set `nAGQ = 0` for the fastest computation because the analysis of these data is solely intended to be illustrative. We store the results as `m0` and use the `tidy` function from `broom.mixed` to print the results in a clean table. While `glmer` models do not provide standard errors or confidence intervals for estimated random effect variances (as indicated by the NA values in the output), users can employ the `confint` function to obtain profile confidence intervals for the variance components of the model, if desired.

```
m0 <- glmer(s_correct ~
1 + (1|s_id) +
(1|item),
data = more_long,
family = binomial,
nAGQ = 0)
tidy(m0)
```

```
## # A tibble: 3 x 7
##   effect   group term                estimate std.error statistic p.value
##   <chr>    <chr> <chr>                <dbl>    <dbl>    <dbl>    <dbl>
## 1 fixed    <NA> (Intercept)         0.239     0.146      1.64     0.102
## 2 ran_pars s_id   sd__(Intercept)     0.945     NA        NA        NA
## 3 ran_pars item  sd__(Intercept)     0.644     NA        NA        NA
```

The model output can be interpreted as follows. The intercept β_0 represents the log-odds of a correct response for the average student to the average item. In probability terms, this translates to 55.95%. σ_e represents the unconditional standard deviation (SD) of the person proficiency parameters, and σ_{ζ_0} represents the unconditional SD of the item easiness parameters, both on the logit scale. In other words,

some students have higher probabilities of a correct response, and some items are easier than others, but we have not (yet) specified any hypotheses as to *why* such differences exist. As such, an EIRM without predictors is known as a “doubly descriptive” model (Wilson & De Boeck, 2004). The EIRM becomes “person explanatory”, “item explanatory”, or “doubly explanatory” when person predictors, item predictors, or both, are added to the model, respectively (ibid).

We can leverage this model to extract the estimated person parameters, representing the estimated proficiency of the students. In a two-step analysis, these estimated scores would then be analyzed as the outcome in a regression model. One benefit of the EIRM is that, as a latent variable model, it performs estimation and regression in a single step, which results in more accurate estimates than two-step approaches (Briggs, 2008; Gilbert, 2022, 2023). The code below uses the `ranef` and `augment` functions to extract the person parameters and plot their density, and we see in Fig. 1 that they are approximately normally distributed.

```
m0 |>
  ranef() |>
  augment() |>
  filter(grp == "s_id") |>
  select(estimate) |>
  ggplot(aes(x = estimate)) +
  geom_histogram() +
  labs(x = "Estimated Theta Score",
       y = "Count")
```

Model 1: Constant treatment effect

We can extend the baseline model by including a treatment effect β_1 and a pretest covariate β_2 in the equation for θ_j :

$$\eta_{ij} = \theta_j + b_i$$

$$\theta_j = \beta_0 + \beta_1 \text{treat}_j + \beta_2 \text{pretest}_j + \varepsilon_j$$

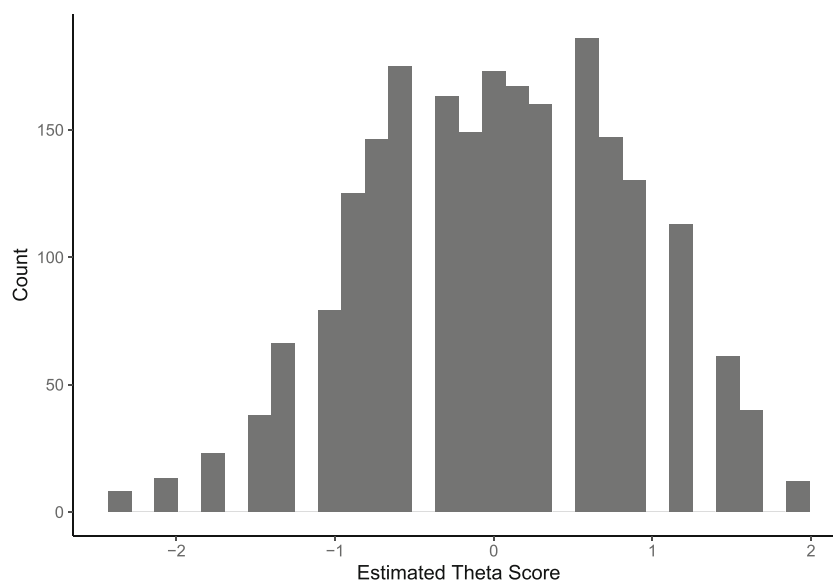
$$b_i = \zeta_{0i}$$

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

$$\varepsilon_j \sim N(0, \sigma_e)$$

In this study, the pretest was the Measure of Academic Progress (MAP) assessment, administered by the state prior to the start of the intervention. We include the pretest variable because it provides substantial precision gains and because the authors reported a slight baseline imbalance between the treatment and control groups for this measure. To fit this model in R, we simply add the `treat` and `pretest` terms

Fig. 1 Distribution of student proficiency estimates derived from Model 0



to the `glmer` argument. We store the result as `m1` and present the results with `tidy`.

```
m1 <- glmer(s_correct ~ treat + pretest
+ (1|s_id) +
(1|item),
more_long,
family = binomial,
nAGQ = 0)
tidy(m1)
```

```
## # A tibble: 5 x 7
##   effect group term          estimate std.error statistic    p.value
##   <chr>   <chr> <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 fixed   <NA> (Intercept)    0.112    0.146      0.764  4.45e- 1
## 2 fixed   <NA> treat         0.186    0.0367     5.08   3.80e- 7
## 3 fixed   <NA> pretest       0.653    0.0193    33.9   6.29e-252
## 4 ran_pars s_id sd__(Intercept) 0.679    NA         NA      NA
## 5 ran_pars item sd__(Intercept) 0.644    NA         NA      NA
```

Here, the intercept β_0 provides the fitted log-odds of a correct response for the average control student with average baseline achievement, to an item of average easiness. The parameter of primary interest is the coefficient for β_1 which represents the ATE on the logit scale, assumed to be constant across items. We see that, controlling for baseline achievement, the MORE treatment caused an increase in θ of about 0.19 logits, and this difference is statistically significant. β_2 shows that for both treatment and control students, a 1SD higher pretest score predicts 0.65 higher logits of a correct response. We can see that the σ_e has decreased from the baseline model by adding person predictors, but σ_{ζ_0} remains unchanged because we have not added any item predictors to the model. σ_e now represents *residual* student proficiency, or differences in student response probabilities not accounted for by pretest scores or treatment status.

Model 2: Random IL-HTE

We extend the model further by allowing for IL-HTE with a random slope for treatment at the item level:

$$\begin{aligned}\eta_{ij} &= \theta_j + b_{ij} \\ \theta_j &= \beta_0 + \beta_1 \text{treat}_j + \beta_2 \text{pretest}_j + \varepsilon_j \\ b_{ij} &= \zeta_{0i} + \zeta_{1i} \text{treat}_j\end{aligned}$$

$$\begin{aligned}\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} &\sim N\left(0, \begin{bmatrix} \sigma_{\zeta_0} & \rho \\ \rho & \sigma_{\zeta_1} \end{bmatrix}\right) \\ \varepsilon_j &\sim N(0, \sigma_e)\end{aligned}$$

In R, we substitute `(treat|item)` for `(1|item)` in the `glmer` argument to allow for IL-HTE and display the results.

```
m2 <- glmer(s_correct ~
treat + pretest +
(1|s_id)
+ (treat|item),
more_long,
family = binomial,
nAGQ = 0)
tidy(m2)
```

```
## # A tibble: 7 x 7
##   effect group term estimate std.error statistic p.value
##   <chr>   <chr> <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 fixed   <NA> (Intercept) 0.114    0.161     0.706 4.80e- 1
## 2 fixed   <NA> treat      0.183    0.0531    3.44 5.75e- 4
## 3 fixed   <NA> pretest    0.653    0.0193   33.9 4.48e-252
## 4 ran_pars s_id sd__(Intercept) 0.679    NA        NA    NA
## 5 ran_pars item sd__(Intercept) 0.711    NA        NA    NA
## 6 ran_pars item cor__(Intercept).treat -0.745    NA        NA    NA
## 7 ran_pars item sd__treat    0.171    NA        NA    NA
```

Before interpreting the parameter estimates, we can test whether the random slope is necessary by conducting a likelihood ratio (LR) test comparing Model 1 and Model 2 with `anova`. We find that the IL-HTE model is a vastly better fit, suggesting that there is significant IL-HTE in the data. This result is even more striking when we consider that the basic LR test provided by `anova` is likely to be conservative because the null hypothesis of 0 variance is on the boundary of the parameter space (i.e., variances cannot be negative), and the true distribution is a mixture distribution, an issue that has received much commentary in multilevel models generally (Molenberghs & Verbeke, 2007; Rabe-Hesketh & Skrondal, 2022; Self & Liang, 1987; Stoel, Garre, Dolan & Wittenboer, 2006; Stram & Lee, 1994). More robust alternatives to the basic LR test include the simulation-based `PBmodcomp` function from the `pbkrtest` package (Halekoh & Højsgaard, 2014) though they are more computationally intensive.

```
anova(m1, m2) |> tidy()
```

```
## # A tibble: 2 x 9
##   term npar AIC BIC logLik deviance statistic df p.value
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 m1 5 52346. 52389. -26168. 52336. NA NA NA
## 2 m2 7 52310. 52371. -26148. 52296. 39.4 2 2.76e-9
```

In the IL-HTE model, β_1 still represents the ATE on the logit scale, but it now represents the treatment effect for the *average* item. We see that the point estimate of β_1 is similar to that of the constant effect model, but the standard error is much larger, a result in line with the simulation study in Gilbert et al. (2023). Conceptually, the larger SE of the IL-HTE model compared to the constant effect random intercepts model reflects the additional variance of the treatment effect estimand in the population of items, in contrast to the finite sample treatment effect on the specific set of items on the realized test (ibid).

The IL-HTE model provides two new parameters. First, σ_{ζ_1} (`sd__treat`) parameterizes the IL-HTE, and estimates the SD of the item-specific treatment effects. We see that the SD is nearly as large as the ATE β_1 , suggesting that the degree of IL-HTE in the data is substantial. Second, ρ (`cor__(Intercept).treat`) represents the strong negative correlation between item easiness and treatment

effect size, suggesting that the easier the item, the lower the magnitude of the treatment effect. In other words, the MORE treatment was most impactful on the most difficult assessment items. ρ can be difficult to interpret in multilevel models generally because a linear transformation of the variable involved in the random slope can change the estimated correlation (Bates, Mächler, Bolker & Walker, 2015). This issue is less relevant in the current case because the binary treatment indicator has a natural zero point. That is, it represents students in the control group.

The item-specific treatment effects are shown in Fig. 2, and we can see the large spread of item effects around the average. The range of item-specific treatment effect sizes is substantial, ranging from slightly negative to about 0.6 logits. Such meaningful variation would be masked by the single point estimate of the constant effect model.

Model 3: Systematic IL-HTE

Model 2 shows that significant IL-HTE exists in this data set and provides a better estimate of the uncertainty of the

ATE than Model 1, but does not test any hypotheses for *why* some items demonstrate larger or smaller effects than others. To test such a hypothesis, we can add item-characteristic by treatment interactions to the model. In other words, we can move from the random IL-HTE of the random slope formulation to systematic IL-HTE of the item-characteristic by treatment interactions. Here, we interact the treatment with the near/mid/far transfer passage indicator to determine whether treatment effect size systematically depends on the transfer passage type:

$$\eta_{ij} = \theta_j + b_i$$

$$\theta_j = \beta_0 + \beta_1 \text{treat}_j + \beta_2 \text{pretest}_j + \varepsilon_j$$

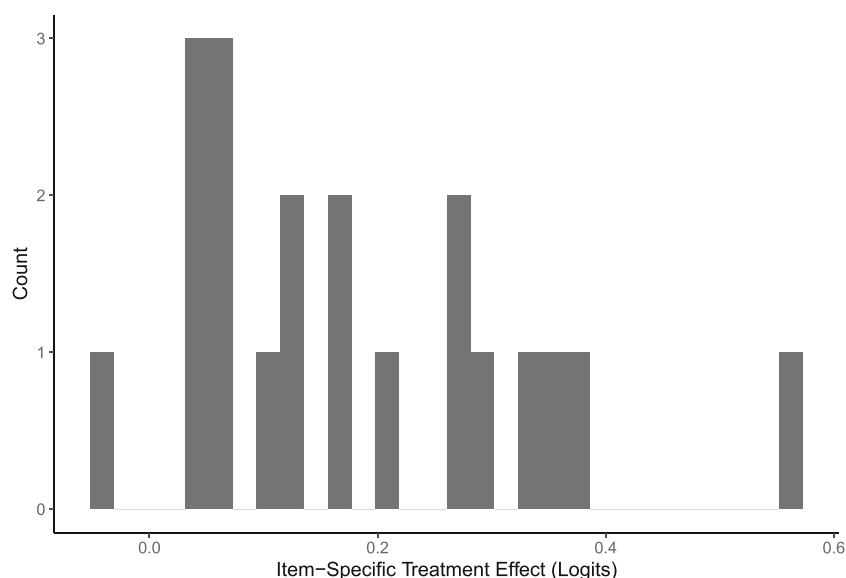
$$b_{ij} = \beta_3 \text{mid}_i + \beta_4 \text{far}_i + \zeta_{0i} + \zeta_{1i} \text{treat}_j$$

$$\beta_{1i} = \gamma_{00} + \gamma_{01} \text{mid}_i + \gamma_{02} \text{far}_i$$

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

$$\varepsilon_j \sim N(0, \sigma_e)$$

Fig. 2 Distribution of item-specific treatment effects derived from Model 2



In R, we use `treat*passage` in the `glmer` argument to create the interaction terms, fit the model, and display the results.

```
m3 <- glmer(s_correct ~
  treat*passage + pretest +
  (1|s_id) + (treat
  |item),
  more_long,
  family = binomial,
  nAGQ = 0)
tidy(m3)
```

```
## # A tibble: 11 x 7
##   effect  group term                estimate std.error statistic    p.value
##   <chr>   <chr> <chr>                <dbl>     <dbl>    <dbl>    <dbl>
## 1 fixed   <NA> (Intercept)        -0.110     0.248    -0.445  6.57e- 1
## 2 fixed   <NA> treat              0.304     0.0691    4.40   1.10e- 5
## 3 fixed   <NA> passagemid         0.0893    0.349     0.256   7.98e- 1
## 4 fixed   <NA> passagefar         0.641     0.364     1.76   7.82e- 2
## 5 fixed   <NA> pretest           0.653     0.0193   33.9   4.31e-252
## 6 fixed   <NA> treat:passagemid  -0.117     0.0872   -1.34   1.80e- 1
## 7 fixed   <NA> treat:passagefar  -0.262     0.0912   -2.87   4.05e- 3
## 8 ran_pars s_id  sd__(Intercept)     0.679     NA        NA      NA
## 9 ran_pars item  sd__(Intercept)     0.650     NA        NA      NA
## 10 ran_pars item  cor__(Intercept).treat -0.698     NA        NA      NA
## 11 ran_pars item  sd__treat           0.132     NA        NA      NA
```

The addition of the transfer passage interaction effects changes the interpretation of the other model coefficients. The coefficient for `treat` now provides the CATE for near transfer items only, as the near transfer passage items are the reference group in this parameterization. The interaction effects `treat:passagemid` and `treat:passagefar` estimate the *difference* in the CATE for those item clusters compared to the CATE for the near transfer items. That is, the CATE for mid transfer items is slightly lower than that

of the near transfer items (though the difference is not statistically significant), and the CATE for far transfer items is significantly lower than that of the near transfer items. This result makes conceptual sense given the structure of the MORE intervention, which would be expected to generate larger effects on items that were closer to the content delivered through MORE. Such a fine-grained finding would potentially be overlooked with other analytic methods, such as a regression model of a sum score or latent variable model that only allows for a single point estimate of the ATE. Furthermore, this result is policy relevant as it provides a direct

measure of “how far intervention effects travel” (Kim et al., 2023), and suggests that the MORE intervention is likely to be most effective on assessments that employ similar content and vocabulary to that of the MORE curriculum.

The main effects of `passagemid` and `passagefar` represent differences in average item easiness compared to the near transfer passage for control students, and we see that they are not significantly different. Furthermore, in this model, we see that σ_{ζ_1} is reduced when compared to the

unconditional IL-HTE, suggesting that the interaction terms explain some of the IL-HTE in the data set.

We can visualize the transfer passage treatment effects using the `ggpredict` function from `ggeffects` (Lüdtke, 2018), which estimates model-predicted values holding the other covariates and random effects constant at their means. Figure 3 shows fitted values on the logit scale for treatment and control students for each transfer passage. Following directly from the model results, we see the largest difference between treatment and control students on the near transfer passage.

```
ggpredict(m3, terms = c("passage", "treat")) |>
  mutate(logit_hat = qlogis(predicted)) |>
  drop_na() |>
  ggplot(aes(x = x, y = logit_hat, shape = group)) +
  geom_point() +
  labs(y = "Log-Odds",
       x = "Transfer Passage Type",
       shape = "Treatment Group") +
  theme(legend.position = "bottom")
```

Finally, for presentation, the results of multiple models can be easily displayed using the `texreg` function, as shown in Table 1.

```
texreg(list(m0, m1, m2, m3),
        single.row = TRUE,
        custom.model.names = c("M0",
                                "M1",
                                "M2",
                                "M3"),
        caption = "Results of Explanatory
                    Item")
```

Response Models fit to the MORE
Intervention Data)

Further explorations

Having built the IL-HTE model to increasing levels of complexity, two additional nuances merit consideration. First, do the transfer passage by treatment interaction terms explain all of the IL-HTE in the data set? Second, is the correlation between item easiness and treatment effect size needed in the model? These questions are easy to address by fitting more constrained models and then comparing them to the results of Model 3 using LR tests.

The code below refits Model 3 omitting the IL-HTE term as `m4`. Model 4 assumes that once the differential treatment effects by transfer passage have been taken into account, there is no additional, unexplained IL-HTE in the data. The LR test shows that Model 3 is strongly preferred, suggesting that significant IL-HTE still remains unexplained in the data set, a function of idiosyncratic characteristics of each item.

```
m4 <- glmer(s_correct ~ treat*
            passage + pretest +
            (1|s_id) + (1|item),
            more_long,
            family = binomial,
            nAGQ = 0)
anova(m3, m4) |> tidy()
```

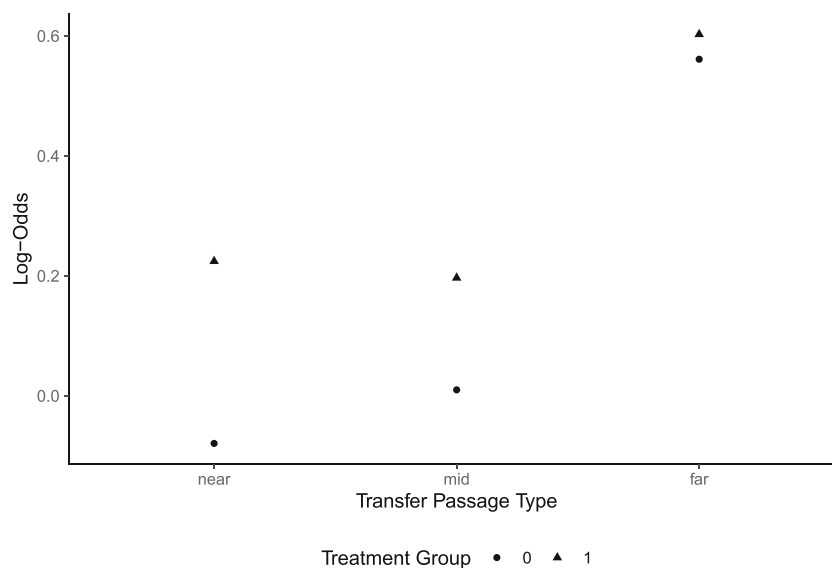


Fig. 3 Model-implied treatment/control contrasts by transfer passage derived from Model 3

Table 1 Results of Explanatory Item Response Models fit to the MORE Intervention Data

	M0	M1	M2	M3
(Intercept)	0.24 (0.15)	0.11 (0.15)	0.11 (0.16)	−0.11 (0.25)
treat		0.19 (0.04)***	0.18 (0.05)***	0.30 (0.07)***
pretest		0.65 (0.02)***	0.65 (0.02)***	0.65 (0.02)***
passagemid				0.09 (0.35)
passagefar				0.64 (0.36)
treat:passagemid				−0.12 (0.09)
treat:passagefar				−0.26 (0.09)**
AIC	53356.40	52345.52	52310.11	52310.17
BIC	53382.44	52388.92	52370.87	52405.65
Log Likelihood	−26675.20	−26167.76	−26148.05	−26144.09
Num. obs.	43480	43480	43480	43480
Num. groups: s_id	2174	2174	2174	2174
Num. groups: item	20	20	20	20
Var: s_id (Intercept)	0.89	0.46	0.46	0.46
Var: item (Intercept)	0.42	0.41	0.51	0.42
Var: item treat			0.03	0.02
Cov: item (Intercept) treat			−0.09	−0.06

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

```
## # A tibble: 2 x 9
##   term      npar    AIC    BIC  logLik deviance statistic    df  p.value
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 m4         9 52326. 52404. -26154.  52308.    NA    NA NA
## 2 m3        11 52310. 52406. -26144.  52288.   19.8     2 0.0000502
```

Similarly, the code below refits Model 3 but specifies the item easiness and item-specific treatment effect size as independent by replacing (treat|item) with (treat||item), with the || indicating independence, and stores the results as m5. Model 5 assumes that the population correlation between item easiness and treatment effect size is 0. The LR test shows again that Model 3 is strongly preferred, suggesting that, within each transfer passage, we observe larger treatment effects on the more difficult items of the assessment.

```
m5 <- glmer(s_correct ~
  treat*passage + pretest +
  (1|s_id) +
  (treat||item),
  more_long,
  family = binomial,
  nAGQ = 0)
ANOVA(m3, m5) |> tidy()
```

```
## # A tibble: 2 x 9
##   term      npar    AIC    BIC  logLik deviance statistic    df  p.value
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 m5        10 52315. 52402. -26148.  52295.    NA    NA NA
## 2 m3        11 52310. 52406. -26144.  52288.   7.21     1 0.00726
```

Extensions

The illustration of the IL-HTE model provided thus far only begins to depict the utility of the EIRM in causal analysis. The models explored in the present study can be easily extended to more diverse assessment contexts, for example by accounting for multiple levels of hierarchy (e.g., students nested within schools), additional covariates, or the simultaneous estimation of person-level and item-level HTE by including treatment by person characteristic interactions in the model. Furthermore, recent advances highlight the power and wide applicability of the EIRM beyond the relatively simple case considered here, which examined only binary responses and the Rasch model in a frequentist framework. First, the EIRM can easily be extended to ordinal responses such as Likert scale items using the `clmm` function (Christensen, 2022) or the `eirm` package (Bulut, Gorgun & Yildirim-Erbaşlı, 2021), or continuous responses using `lmer`. Second, the Rasch or IPL formulation employed in this study assumes that all

items are equally discriminating with respect to the latent trait θ , which is unlikely to be true in practice. While 2PL models are not estimable with lme4 (unless item discrimination parameters are known in advance), the R package PLmixed can estimate 2PL EIRMs (Jeon & Rockwood, 2017). However, PLmixed is not directly applicable to IL-HTE analysis because it does not allow the user to specify the items as random. Furthermore, 2PL EIRMs estimated with PLmixed may suffer from accuracy issues (Zhang, Ackerman & Wang, 2021), and simulations in Gilbert et al. (2023) showed that the performance of the 1PL EIRM is quite robust to a 2PL data generating process. If a 2PL model for IL-HTE were desired, a frequentist approach using Mplus (Petscher, Compton, Steacy & Kinnon, 2020) or a Bayesian approach using the R package brms are viable options (Bürkner, 2021).

To provide just one example of the many possibilities available to researchers when modeling IL-HTE, we refit Model 2 in a Bayesian framework using brms and allow for randomly varying item discriminations in a 2PL model (readers are directed to Bürkner (2021) for full tutorial on fitting IRT models and the EIRM in brms). Based on the results above from the 1PL analysis we set moderately informative priors, with normal distributions for fixed effects and half normal distributions for random effects. The $|i|$ notation allows the item easiness and discrimination parameters to be correlated. Note that the estimation of brms models is computationally intensive (taking about 90 min on the author's personal computer due to the Markov Chain Monte Carlo procedure, in contrast to a few seconds for the simpler 1PL glmer models), and σ_e must be set to 1 for model identification (whereas it was freely estimated in the 1PL model), so the magnitudes of the point estimates cannot be directly compared between the glmer and brms output. The results tell the same essential story as the 1PL analysis, showing substantial IL-HTE in the sd_eta_treat parameter. We also observe substantial variability in item discriminations as shown by the $sd_logalpha_(\text{Intercept})$ parameter, suggesting that a 2PL model may provide a more realistic model for this data. Figure 4 shows the estimated discrimination parameters for each item with 95% credible intervals.

```
## # A tibble: 11 x 8
```

##	effect	component	group	term	estimate	std.error	conf.low	conf.high
##	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
##	1 fixed	cond	<NA>	eta_(Intercep~	0.0137	0.257	-0.486	0.552
##	2 fixed	cond	<NA>	eta_treat	0.315	0.0801	0.154	0.469
##	3 fixed	cond	<NA>	eta_pretest	0.987	0.0345	0.920	1.05
##	4 fixed	cond	<NA>	logalpha_(Int~	-0.363	0.106	-0.572	-0.148
##	5 ran_pars	cond	item	sd_eta_(Inte~	1.17	0.208	0.836	1.66
##	6 ran_pars	cond	item	sd_eta_treat	0.271	0.0671	0.157	0.424
##	7 ran_pars	cond	item	sd_logalpha_~	0.464	0.0812	0.335	0.652
##	8 ran_pars	cond	s_id	sd_eta_(Inte~	1	0	1	1
##	9 ran_pars	cond	item	cor_eta_(Int~	-0.617	0.187	-0.894	-0.179
##	10 ran_pars	cond	item	cor_eta_(Int~	0.411	0.192	-0.0192	0.729
##	11 ran_pars	cond	item	cor_eta_trea~	-0.431	0.217	-0.793	0.0512

```
formula_2pl <- bf(
  s_correct ~ exp(logalpha)*
  eta, # model for response
  eta ~ 1 + treat +
  pretest + (treat|i|
  item) + (1|s_id),
  # model for eta
  logalpha ~ 1 +
  (1|i|item),
  # model for discrimination
  nl = TRUE
  # declare non-linear model)
  # set priors
prior_2pl <-
  prior("normal(0.1, 0.5)", class =
  "b", coef = "treat", nlpar =
  "eta") +
  prior("normal(1, 0.2)", class = "b"
  ,coef = "pretest", nlpar =
  "eta") +
  prior("constant(1)", class = "sd"
  , group = "s_id", nlpar =
  "eta") +
  prior("normal(0, 2)", class = "sd"
  , group = "item", nlpar =
  "eta") +
  prior("normal(0, 1)", class =
  "sd", group = "item", nlpar = "logalpha"
  )# fit the model fit_2pl <-brm(
  formula = formula_2pl,
  data = more_long,
  family = brmsfamily("bernoulli",
  "logit"),prior = prior_2pl,
  chains = 4,
  iter = 2000)

# display the results
tidy(fit_2pl)
```

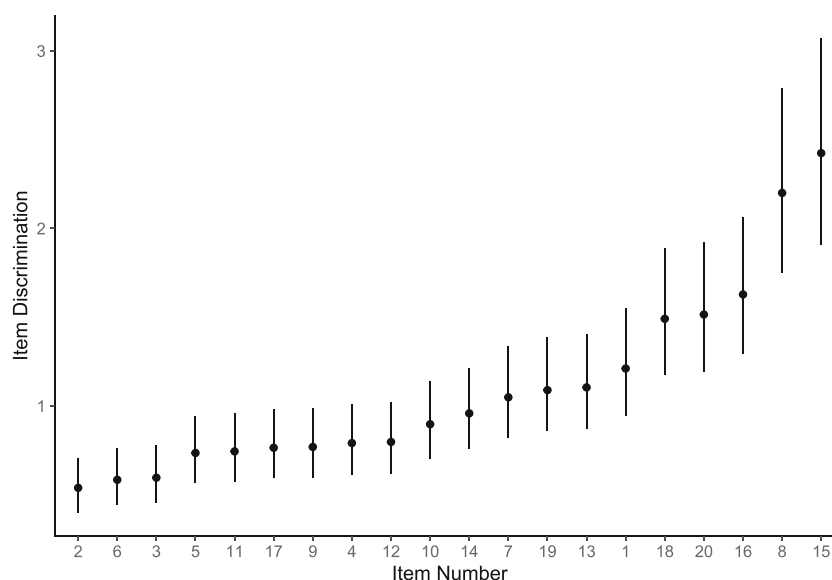


Fig. 4 Estimated Item Discriminations Derived from brms

Conclusions

The IL-HTE model represents an advance in causal analysis, enabling researchers to move beyond traditional person-focused approaches and delve into the heterogeneous treatment effects that may exist within an outcome measure at the item level. By incorporating this innovative method into their analysis, researchers can generate new insights, leading to more informed decision-making and targeted interventions in various domains. By employing the IL-HTE model, researchers gain both statistical and substantive insights, which contribute to a deeper understanding of treatment effects. The worked example utilizing the MORE data set demonstrates the potential of the IL-HTE model in providing fine-grained and meaningful results that would otherwise remain hidden when relying solely on a single summary score analysis. Furthermore, this paper provides a comprehensive tutorial that equips researchers with the necessary tools to successfully apply and adapt the IL-HTE model to their own data sets, in both simple and complex circumstances. By following the step-by-step guidelines and leveraging the provided resources of this tutorial, researchers can harness the power of the IL-HTE model to enhance their analyses and generate more nuanced and meaningful findings.

Funding No financial support was received for the writing of this article.

Availability of data and materials The code for this article (which provides access to the already public data) is available as an online supplemental file.

Code Availability The code for this article is available as an online supplemental file (<https://researchbox.org/2054>).

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflicts of interest The author reports no conflicts of interest.

References

- Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, 21(1), 1–23. <https://doi.org/10.1177/0146621697211001>
- Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., & Domingue, B. W. (2023). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials. *Edworkingpapers.com*. <https://doi.org/10.26300/1NW4-NA96>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Briggs, D. C. (2008). Using Explanatory Item Response Models to Analyze Group Differences in Science Achievement. *Applied Measurement in Education*, 21(2), 89–118. <https://doi.org/10.1080/08957340801926086>
- Bulut, O., Gorgun, G., & Yildirim-Erbaşlı, S. N. (2021). Estimating Explanatory Extensions of Dichotomous and Polytomous Rasch

- Models: The *lme4* Package in R. *Psych*, 3(3), 308–321. <https://doi.org/10.3390/psych3030023>
- Bürkner, P.-C. (2021). Bayesian Item Response Modeling in R with **brms** and *Stan*. *Journal of Statistical Software*, 100(5). <https://doi.org/10.18637/jss.v100.i05>
- Christensen, R. H. B. (2022). *Ordinal—regression models for ordinal data*.
- De Boeck, P. (2008). Random Item IRT Models. *Psychometrika*, 73(4), 533–559. <https://doi.org/10.1007/s11336-008-9092-x>
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The Estimation of Item Response Models with *thelme4* Function from the **thelme4** Package in R. *Journal of Statistical Software*, 39(12). <https://doi.org/10.18637/jss.v039.i12>
- Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the Multilevel Rasch Model: With the **lme4** Package. *Journal of Statistical Software*, 20(2). <https://doi.org/10.18637/jss.v020.i02>
- Francis, D. J., Kulesz, P. A., Khalaf, S., Walczak, M., & Vaughn, S. R. (2022). Is the treatment weak or the test insensitive: Interrogating item difficulties to elucidate the nature of reading intervention effects. *Learning and Individual Differences*, 97, 102167. <https://doi.org/10.1016/j.lindif.2022.102167>
- Gilbert, J. B. (2022). Estimating treatment effects with the explanatory item response model. *EdWorkingPapers.com*. <https://doi.org/10.26300/SNVZ-EW19>
- Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling Item-Level Heterogeneous Treatment Effects With the Explanatory Item Response Model: Leveraging Large-Scale Online Assessments to Pinpoint the Impact of Educational Interventions. *Journal of Educational and Behavioral Statistics*, 107699862311717. <https://doi.org/10.3102/10769986231171710>
- Gilbert, Joshua B. (2023). How measurement affects causal inference: Attenuation bias is (usually) more important than scoring weights. *Edworkingpapers.com*. <https://doi.org/10.26300/4HAH-6S55>
- Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package **pbkrtest**. *Journal of Statistical Software*, 59(9). <https://doi.org/10.18637/jss.v059.i09>
- Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, 6(2), 107–128. <https://doi.org/10.3102/10769986006002107>
- Jeon, M., & Rockwood, N. (2017). PLmixed: An R Package for Generalized Linear Mixed Models With Factor Structures. *Applied Psychological Measurement*, 42(5), 401–402. <https://doi.org/10.1177/0146621617748326>
- Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., & McIntyre, J. (2023). 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*, 115(1), 73–98. <https://doi.org/10.1037/edu0000751>
- Koretz, D. (2005). Alignment, High Stakes, and the Inflation of Test Scores. *Yearbook of the National Society for the Study of Education*, 104(2), 99–118. <https://doi.org/10.1111/j.1744-7984.2005.00027.x>
- Lüdtke, D. (2018). Ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, 3(26), 772. <https://doi.org/10.21105/joss.00772>
- Molenberghs, G., & Verbeke, G. (2007). Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space. *The American Statistician*, 61(1), 22–27. <https://doi.org/10.1198/000313007x171322>
- Montoya, A. K., & Jeon, M. (2019). MIMIC Models for Uniform and Nonuniform DIF as Moderated Mediation Models. *Applied Psychological Measurement*, 44(2), 118–136. <https://doi.org/10.1177/0146621619835496>
- Naumann, A., Hochweber, J., & Hartig, J. (2014). Modeling Instructional Sensitivity Using a Longitudinal Multilevel Differential Item Functioning Approach. *Journal of Educational Measurement*, 51(4), 381–399. <https://doi.org/10.1111/jedm.12051>
- 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. <https://doi.org/10.1007/s11881-020-00204-y>
- Rabe-Hesketh, S., & Skrondal, A. (2022). *Multilevel and longitudinal modeling using stata*. STATA press.
- Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8(2), 185–205. <https://doi.org/10.1037/1082-989x.8.2.185>
- Sales, A., Prihar, E., Heffernan, N., & Pane, J. F. (2021). The effect of an intelligent tutor on performance on specific posttest problems. *International Educational Data Mining Society*.
- Self, S. G., & Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Non-standard Conditions. *Journal of the American Statistical Association*, 82(398), 605–610. <https://doi.org/10.1080/01621459.1987.10478472>
- Stoel, R. D., Garre, F. G., Dolan, C., & Wittenboer, G. van den. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, 11(4), 439–455. <https://doi.org/10.1037/1082-989x.11.4.439>
- Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4), 1171. <https://doi.org/10.2307/2533455>
- Wilson, M., & De Boeck, P. (2004). *Descriptive and explanatory item response models*. Springer New York. https://doi.org/10.1007/978-1-4757-3990-9_2
- Zhang, J., Ackerman, T., & Wang, Y. (2021). *2PL model: Compare generalized linear mixed model with latent variable model based IRT framework*. Retrieved from <https://doi.org/10.31234/osf.io/p6wuz>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.