EdWorkingPaper No. 23-868

Leveraging Item Parameter Drift to Assess Transfer Effects in Vocabulary Learning

Joshua B. Gilbert
Harvard University

James S. Kim
Harvard University

Luke W. Miratrix
Harvard University

Longitudinal models of individual growth typically emphasize between-person predictors of change but ignore how growth may vary within persons because each person contributes only one point at each time to the model. In contrast, modeling growth with multi-item assessments allows evaluation of how relative item performance may shift over time. While traditionally viewed as a nuisance under the label of "item parameter drift" (IPD) in the Item Response Theory literature, we argue that IPD may be of substantive interest if it reflects how learning manifests on different items or subscales at different rates. In this study, we present a novel application of the Explanatory Item Response Model (EIRM) to assess IPD in a causal inference context. Simulation results show that when IPD is not accounted for, both parameter estimates and their standard errors can be affected. We illustrate with an empirical application to the persistence of transfer effects from a content literacy intervention on vocabulary knowledge, revealing how researchers can leverage IPD to achieve a more fine-grained understanding of how vocabulary learning develops over time.

VERSION: May 2024

Suggested citation: Gilbert, Joshua B., James S. Kim, and Luke W. Miratrix. (2024). Leveraging Item Parameter Drift to Assess Transfer Effects in Vocabulary Learning. (EdWorkingPaper: 23-868). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/hpzb-2f41

Leveraging Item Parameter Drift to Assess Transfer Effects in Vocabulary Learning

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

Harvard University Graduate School of Education

Corresponding Author: joshua_gilbert@g.harvard.edu

Forthcoming in *Applied Measurement in Education*

Abstract

Longitudinal models of individual growth typically emphasize between-person predictors of change but ignore how growth may vary *within* persons because each person contributes only one point at each time to the model. In contrast, modeling growth with multi-item assessments allows evaluation of how relative item performance may shift over time. While traditionally viewed as a nuisance under the label of "item parameter drift" (IPD) in the Item Response Theory literature, we argue that IPD may be of substantive interest if it reflects how learning manifests on different items or subscales at different rates. In this study, we present a novel application of the Explanatory Item Response Model (EIRM) to assess IPD in a causal inference context. Simulation results show that when IPD is not accounted for, both parameter estimates and their standard errors can be affected. We illustrate with an empirical application to the persistence of transfer effects from a content literacy intervention on vocabulary knowledge, revealing how researchers can leverage IPD to achieve a more fine-grained understanding of how vocabulary learning develops over time.

Keywords: Latent Growth Curve, Explanatory Item Response Model, Causal Inference, Simulation, Psychometrics

Leveraging Item Parameter Drift to Assess Transfer Effects in Vocabulary Learning

Longitudinal models of individual growth provide critical insight into the developmental patterns of learning in educational research. The multilevel modeling literature in particular highlights the flexibility and power of longitudinal models to assess average growth trends across a population, heterogeneity in individual trajectories, and systematic predictors of variation in individual growth (Singer & Willett, 2003). Modeling individual growth heterogeneity is particularly important because it enables researchers to both quantify how much heterogeneity in growth exists in a population and simultaneously model the systematic sources of that heterogeneity. Paraphrasing Raudenbush and Bloom (2015), assessing individual growth heterogeneity through random slope models allows researchers to learn *about* individual growth heterogeneity, whereas subject characteristic by time interactions allow researchers to learn *from* individual growth heterogeneity.

In longitudinal studies that examine repeated measures on a single outcome measure over time, only between-subject predictors (or predictors at higher levels of clustering such as schools) of individual growth heterogeneity can be evaluated because each subject provides only one data point per measurement occasion, and the model is therefore unable to assess potential within-subject heterogeneity in growth profiles among different items or subscales of an outcome measure. This limitation can be overcome, however, by extending the longitudinal model to incorporate item-level data from an assessment, rather than a single summary score representing that assessment, because each subject contributes multiple data points per measurement occasion (Luo, et al., 2022; Marvelde, et al., 2006). Just as persons may vary randomly in their growth trajectories and person characteristic (e.g., age, gender, demographic group) by time interactions can explain average between-*person* differences in growth rates, proficiency on items may vary randomly in their growth trajectories and item characteristic (e.g., subscale, modality, content area) by time interactions can explain between-*item* differences in growth rates. For example, a longitudinal model of vocabulary items could reveal whether treatment effects on growth rates are greater on explicitly taught vocabulary words compared to untaught vocabulary words, a possibility we examine in our empirical application.

Longitudinal models of item-level data are also known as latent growth curve models (LGCM) or longitudinal item response theory (LIRT) models because the item responses can be interpreted as indicators of a latent construct that develops over time, such as vocabulary or reading comprehension ability. There is a rich body of literature on the utility and application of LIRT and associated models. These models have been used to simultaneously assess average change in a latent trait, between-person heterogeneity in that change, and psychometric properties such as longitudinal measurement invariance (Pastor & Beretvas, 2006), and can be extended to higher-order latent variable structures (Wang & Nydick, 2019), polynomial growth trajectories, differential item functioning (DIF), serial dependence or autocorrelation (Jeon & Rabe-Hesketh, 2016), and multidimensionality (Wilson, et al., 2012; Cho et al., 2013). Prior simulation studies have demonstrated that Generalized Linear Mixed Model (GLMM) or Explanatory Item Response Model (EIRM) estimation procedures were generally superior to other methods such as two-step approaches, structural equation modeling (SEM), and Bayesian Markov Chain Monte Carlo (MCMC) methods (Ye, 2016). Empirical causal inference applications are rare but include Stevenson and colleagues (2013), who applied a longitudinal EIRM in a pre-post design measuring students' change on a test of analogical reasoning, in which differential student growth rates were modeled as a function of the randomly assigned treatment.

One challenge of LIRT and associated models is the possibility of failures of longitudinal measurement invariance (LMI), also called item parameter drift (IPD; Rupp & Zumbo, 2006). That is, item discriminations (i.e., factor loadings), item difficulties (i.e., item intercepts), and factor variances may shift over time. LMI/IPD exists in three forms: (a) "weak" invariance, in which the item discriminations are equal over time, (b) "strict" invariance, in which the discriminations and difficulties are equal, and (c) "strong" invariance, in which discrimination, difficulties, and factor variances are equal (Liu, et al., 2017). Luo and colleagues (2020) argued that "LMI is a desirable quality in a measurement because it indicates that the same construct can be tested across occasions … providing a solid and necessary basis for mean comparisons in longitudinal studies. Any inference about developmental changes over time may be misleading and inaccurate unless the premise of LMI is met" (pp. 2-3). Various methods exist for detecting violations of LMI, such as Lord's χ^2 (Donoghue & Isham, 1998), and when it is ignored, parameter bias, inaccurate confidence intervals, and scoring inaccuracy may result (Lee & Cho, 2017; Lee & Geisinger, 2019).

A commonality in the LIRT literature is that LMI or IPD is treated as a nuisance to be evaluated and addressed rather than as an object of substantive interest. That is, changes in item parameters over time may cast doubt on the validity of the construct being longitudinally measured, and items exhibiting drift may need to be removed from the data to better meet the assumptions of the model. In contrast, Sukin (2010) argued that IPD may reflect a substantively meaningful pattern of differential learning across different types of items, noting that "if the performance of the item has *changed due to improvements in instruction*, then removing the anchor item [i.e., the item exhibiting IPD] may not be appropriate and might produce misleading conclusions about the proficiency of the examinees" (p. vii, emphasis added). Similarly, VanderWeele and Vansteelant (2022) argued that the indicators of a latent factor (i.e., the items) may themselves be of causal interest beyond their role in measurement of the latent factor, even when assumptions of unidimensionality are met, because of the unique information provided by each indicator (see also MIMIC models, e.g., Montoya & Jeon, 2020).

In this study, we argue that explicit modeling of IPD using random slopes models that allow for variation of item growth rates around the average growth rate and interpreting it as a substantively interesting feature of longitudinal data capable of revealing more fine-grained profiles of student development can provide opportunities to better understand individual growth. By modeling IPD, researchers can go beyond average between-person trends and better understand how, within students, proficiency on different individual items or subscales develops. When modeled appropriately, IPD may provide new insights into longitudinal growth in both descriptive and causal inference contexts. Accordingly, we propose a novel application of the Explanatory Item Response Model (EIRM) to quantify and leverage IPD for deeper understanding of individual growth. We first use simulation to test the performance of the EIRM with and without IPD in a causal inference context. We then apply the EIRM to empirical three-year longitudinal vocabulary assessment data from the Model of Reading Engagement (MORE) content literacy intervention for early elementary grades (see Kim, et al., 2021, 2023, 2024 for prior studies of MORE).

The Explanatory Item Response Model (EIRM)

In its simplest form, and without a longitudinal structure, the Explanatory Item Response Model (EIRM; De Boeck, et al., 2016) is a cross-classified logistic regression model, in which responses are nested within the cross-classification of items and persons. Consider the following model,

$$logit\left(P(y_{ij} = 1)\right) = \beta_0 + \theta_{0j} + \zeta_{0i} \tag{1}$$

$$\theta_{0j} \sim N(0, \sigma^2_{\theta_0})$$

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

in which the log-odds of a correct response of person j to dichotomous item i is a function of a constant term (β_0), person ability (θ_{0j}), and item easiness (ζ_{0i}). Person ability and item easiness are assumed to be normally distributed with mean 0 and some variance ($\sigma^2_{\theta_0}$ and $\sigma^2_{\zeta_0}$ respectively) for model identification. Persons and items can be modeled as either fixed or random effects, or a combination of the two, but persons are almost always modeled as random (De Boeck, 2008). When persons are random, items are fixed, and there are no predictors in the model, the EIRM is mathematically equivalent to a Rasch or 1PL IRT model. Building on prior studies employing the EIRM (Gilbert, Kim, & Miratrix, 2023), here, we consider the random effect specification for the items because (a) it treats items as a source of variability, an approach that is conceptually appropriate when inference to the population of items are of interest (such as when items are drawn from a pool of potential items), (b) the standard errors for the fixed effects in the model reflect the sampling error of which items were selected for test administration (i.e., in contrast to the finite sample, test-specific estimand; see ibid; Miratrix, et al., 2021, for a discussion), and (c) it provides the ability to model random slopes for time at the item level to evaluate IPD, a possibility we now explore adding a longitudinal dimension to the model.

We can extend the cross-sectional EIRM to longitudinal contexts (Cho, et al., 2013; Wilson, et al., 2012) with a linear growth EIRM by adding the subscript *t* to indicate measurement occasions across time, a fixed effect for time to capture the average growth rate, and a random slope for time at the person level:

Random Slopes for Persons, Random Intercepts for Items (2)

$$logit\left(P(y_{tij} = 1)\right) = \beta_0 + \beta_1 time_{tij} + \theta_{0j} + \zeta_{0i} + \theta_{1j} time_{tij}$$

$$\begin{bmatrix}\theta_{0j}\\ \theta_{1j}\end{bmatrix} \sim N(0, \begin{bmatrix}\sigma^2_{\theta_0} & \rho_{10}\\ \rho_{01} & \sigma^2_{\theta_1}\end{bmatrix})$$

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

Here, β_0 is the log-odds of a correct response at baseline (time = 0), and β_1 is the linear growth rate in the log-odds of a correct response over time, averaged across students and items. Conceptually, β_1 represents changing proficiency or ability over time. The linear functional form of the growth rate can easily be extended to polynomial or piecewise specifications if desired. The random intercept term θ_{0j} represents the deviation of person ability from the average ability at baseline β_0 and the random slope term θ_{1j} represents the deviation of each person's growth rate from the average growth rate β_1, averaged across items, with mean 0 and variances $\sigma^2_{\theta_0}$ and $\sigma^2_{\theta_1}$, respectively. ζ_{0i} represents item easiness and is assumed to be constant across time with mean 0 and variance $\sigma^2_{\zeta_0}$. ρ_{01} represents the correlation between random intercepts and random slopes, and would reveal whether, for example, initially high-achieving students demonstrated lower or higher growth rates than initially low-achieving students, on average.

The contribution of this study is to explore the consequences of extending the random slope specification simultaneously to the *item* side of the EIRM to represent IPD, in which ζ_{1i} represents the deviation of each item's growth rate from the average growth rate β_1, averaged across persons, and ζ_{0i} now represents item easiness at baseline:

Random Slopes for Persons and Items

$$logit\left(P(y_{tij} = 1)\right) = \beta_0 + \beta_1 time_{tij} + \theta_{0j} + \zeta_{0i} + \theta_{1j} time_{tij} + \zeta_{1i} time_{tij} \quad 3$$

$$\begin{bmatrix}\theta_{0j}\\ \theta_{1j}\end{bmatrix} \sim N(0, \begin{bmatrix}\sigma^2_{\theta_0} & \rho_{10}\\ \rho_{01} & \sigma^2_{\theta_1}\end{bmatrix})$$

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

Substantively, the random slope for time at the item level would indicate that, averaged across persons, proficiency on individual items grows at a unique rate. The variance of item specific growth rates, represented by the parameter $\sigma^2_{\zeta_1}$, implies that the relative item easiness parameters are *not* necessarily fixed over time as they were in the item random intercept models, but rather "drift" at a unique rate for each item. That is, a non-zero estimate of $\sigma^2_{\zeta_1}$ indicates the presence of IPD in the data, such that item i has a growth rate of $\beta_1 + \zeta_{1i}$, and the variance of item-specific growth rates around the average growth rate β_1 is captured by $\sigma^2_{\zeta_1}$.

In other words, $\sigma^2_{\zeta_1}$ provides the total amount of IPD in the data and the residual IPD for any specific item ζ_{1i} can be calculated after the model is fit, for example using empirical Bayes estimation procedures (Ten Have & Localio, 1999; Waclawiw & Liang, 1994; Liu, Kuppens & Bringmann, 2021). The mean of ζ_{1i} is constrained to 0 for model identification because a non-zero mean for ζ_{1i} would be confounded with an equivalent change to the average growth rate β_1. The parameterization of IPD as a random effects variance is analogous to random effects differential item functioning (DIF) approaches in cross-sectional contexts that assume the overall measure is free of DIF but allow each item parameter to deviate randomly about the average (Van den Noortgate & De Boeck, 2005; Frederickx, et al., 2010; Binici, 2007; Gamerman, Goncalves, & Soares, 2017; Gilbert, Kim, & Miratrix, 2023). τ_{01} captures potential associations between item easiness at baseline and item growth rate. For example, a positive value of τ_{01} would suggest that items that were easier at baseline demonstrated higher growth rates, on average, analogous to the parameter ρ_{01} for the person random effects[1].

As discussed earlier, differential growth by item is typically either ignored or interpreted as a nuisance, under the label of IPD, uniform DIF (De Boeck, et al., 2011, pp. 18-19; Randall, Cheong, & Engelhard, 2011), or violations of assumptions of longitudinal measurement invariance, and various strategies have been proposed to detect and adjust for it, including IRT and SEM-based approaches (Lee & Cho, 2017; Proust-Lima et al., 2021). However, we argue that rather than a nuisance, IPD, represented either by the random variation of item growth rates (i.e., $\sigma^2_{\zeta_1}$), growth trajectories for individual items (i.e., $\beta_1 + \zeta_{1i}$), or systematic variation that interacts time with item features that we explore next (e.g., time by subscale interactions) estimated with the EIRM may provide substantive insight into student learning processes in both descriptive and causal contexts, as students' performance on different items or subscales may truly develop at different rates, rather than representing an unreliable or defective assessment instrument. Under this parameterization, therefore, each person-item combination has a unique growth trajectory, and heterogeneous growth can occur both *between* persons (within items, i.e., $\beta_1 + \theta_{1j}$) and *within* persons (between items, i.e., $\beta_1 + \zeta_{1i}$). As such, modeling IPD with the EIRM can allow a more fine-grained insight into the nature of longitudinal growth than the comparable model with item random intercepts alone, or longitudinal models of total test scores.

In addition to allowing quantification of IPD with a random slope for time at the item level, the EIRM can be extended with additional fixed effects at the person- or item-level to answer substantive research questions. For example, in a causal inference context, it is possible to include

[1] While outside the scope of this study, prior research has demonstrated that misspecification of the correlation between random effects can create bias in estimated interactions among the fixed effects. See Gilbert, Miratrix, et al. (2024) for a full treatment of this issue.

a person-level treatment variable and its interaction with time to determine if treatment causes an increase in average growth rates. Similarly, item-level predictors such as subscale (e.g., taught vs. untaught vocabulary words) interacted with time could provide substantive insight into systematic variation in item growth rates, and three-way interactions between treatment, subscale, and time would reveal the extent to which types of items are most benefited by treatment over time, or what has been referred to as "instructional sensitivity" in descriptive contexts (Naumann, Hochweber, & Hartig, 2014), and "item-level heterogeneous treatment effects" (Gilbert, Kim, & Miratrix, 2023) or "item-treatment interactions" (Ahmed, et al., 2023) in causal contexts. For example, consider the following model, where $treat_j$ is an indicator for treatment status and $itemtype_i$ is an indicator for which subscale an item belongs to (e.g., taught vs. untaught vocabulary words):

Random Slopes for Persons and Items with Treatment Effects on Two Subscales

(Varying Person and Item Growth)

$$
\begin{aligned}
logit\left(P\left(y_{tij}=1\right)\right) & \\
&= \beta_0 + \beta_1 time_{tij} + \beta_2 treat_j + \beta_3 treat \times time_{tij} \\
&+ \beta_4 itemtype_i + \beta_5 treat \times itemtype_{ij} \\
&+ \beta_6 itemtype \times time_{ti} + \beta_7 itemtype \times treat \times time_{tij} \\
&+ \theta_{0j} + \zeta_{0i} + \theta_{1j} time_{tij} + \zeta_{1i} time_{tij}
\end{aligned} \tag{4}
$$

$$
\begin{bmatrix}\theta_{0j}\\ \theta_{1j}\end{bmatrix} \sim N(0, \begin{bmatrix}\sigma^2_{\theta_0} & \rho_{10}\\ \rho_{01} & \sigma^2_{\theta_1}\end{bmatrix})
$$

$$
\begin{bmatrix}\zeta_{0j}\\ \zeta_{1j}\end{bmatrix} \sim N(0, \begin{bmatrix}\sigma^2_{\zeta_0} & \tau_{10}\\ \tau_{01} & \sigma^2_{\zeta_1}\end{bmatrix}).
$$

To aid in the interpretation of the many model parameters above, as a concrete example, consider an intervention measured longitudinally with a vocabulary test that includes both words explicitly taught through the intervention and untaught words, with untaught words as the reference category (i.e., $itemtype_i = 1$ for the taught words subscale), as is the case in our empirical application. In this case, the model allows for a treatment-control difference at baseline for untaught words (β_2), a two-way treatment × time interaction (β_3) to determine whether treatment students improved more over time than control students for untaught words (averaged across items), a main effect for differences in easiness between taught and untaught words for control students at baseline (β_4), a treatment-control difference at baseline for taught words (β_5, above any difference on untaught words), a two-way item-type × time interaction to determine whether students demonstrated more growth in taught or untaught words (averaged across persons) (β_6), and a three-way interaction treatment × item-type × time to determine whether treatment effects on growth rates differ by whether the word was taught or untaught, thus revealing within-outcome heterogeneous treatment effects (β_7). A random slope term could also be added for the treatment × time interaction at the item level to represent residual item-level heterogeneous treatment effects if desired (see Gilbert, Kim, & Miratrix, 2023, for a detailed review of modeling item-level heterogeneous treatment effects with the EIRM in cross-sectional contexts and Gilbert, 2023 for a tutorial on fitting such models in R).

By including interaction effects in Equation 4 (i.e., $\beta_5, \beta_6, \beta_7$), $\sigma^2_{\zeta_1}$ now represents the *residual* IPD variance, or the IPD that is not accounted for by any systematic item-type by time interactions. While estimates of item-specific growth rates from models without interactions (i.e., $\beta_1 + \zeta_i$ derived from Equation 3) may be meaningful and useful in exploratory analyses (Sales, et al., 2021), we view the most informative applications of our proposed approach to be the estimation of interactions between item features and time, treatment status, or both, because such interactions can reveal on what *types* of items growth rates may systematically differ, above and beyond any idiosyncratic growth rate for an individual item.

As an analogy, consider a traditional longitudinal growth model of persons. While examining person-specific trajectories can be useful in, say, assessing model fit or potential outliers, person-characteristic by time interactions are typically more meaningful to the researcher because they allow for specific hypothesis tests such as whether growth rates differ systematically by, for example, age, demographic group, gender, socio-economic status, or other person characteristics. We can apply similar reasoning to the item case through interaction effects including item characteristics. For example, prior EIRM applications examine differential treatment by subscale effects in cross-sectional contexts, such as treatment by reading passage interactions in literacy interventions (Kim, et al., 2023; Gilbert, Kim, & Miratrix, 2023; Gilbert, 2023) and treatment by subscale interactions in clinical trials of patient reported depression surveys (Gilbert, Hieronymus, et al., 2024), findings with important policy implications that would not be readily apparent from a model of random item effects alone and no treatment by item characteristic interactions. In other words, the total amount of IPD given by $\sigma^2_{\zeta_1}$ in an unconditional model tells us *that* items vary in their growth rates, but not *why* items vary; understanding why items vary in their growth rates is an insight that only item-type by time interactions (e.g., β_6) can provide (Raudenbush & Bloom, 2015). Conversely, if $\sigma^2_{\zeta_1}$ is 0, there is no need to test for interaction effects because there is no IPD variance to explain. Finally, we can compare unconditional models to interaction models and examine the change in $\sigma^2_{\zeta_1}$ to estimate the proportion of IPD variance explained by the interaction effects as a metric for how well our model has explained the observed IPD variance with systematic item features (Hox, et al, 2017).

Monte Carlo Simulation

We use Monte Carlo simulations conducted in R (R Core Team, 2022) to test the performance of the EIRM with and without IPD. Following previous simulation studies on longitudinal item response models (e.g., Lee & Cho, 2017), and to maintain focus on the effects of IPD, we use Equation 4 as our data-generating model and we fix the number of subjects at 500, the number of items at 20 (representing two subscales of 10 items each, e.g., taught vs. untaught words in our hypothetical example above and our empirical application), the number of repeated measurements at 5, the average growth rate for control students (β_1) at 0.20 logits, the average treatment effect on the reference subscale (β_3, e.g., untaught words) at 0.20 logits, the standard deviation of item easiness (σ_{ζ_0}) and person ability (σ_{θ_0}) at baseline at 1, and the standard deviation of person growth rates (σ_{θ_1}) at 0.10 logits. We explore the combination of two varying factors, the average treatment effect on the focal subscale (β_7, e.g., taught words) at 0 and 0.20 logits and the standard deviation of item level growth rates (σ_{ζ_1}) at 0, 0.10, and 0.20 logits. Thus, we employ a 2×3 factorial design with null and positive subscale interaction effects at no, moderate, and high IPD. When IPD is positive, we have violated "strict" longitudinal measurement invariance, as the item discriminations are constant across time (i.e., our data generating process is a 1PL or Rasch model), but the difficulties are not. The IPD random slopes (σ_{ζ_1}, and the person random slopes, σ_{θ_1}) also imply a violation of "strong" longitudinal measurement invariance because of the heteroskedasticity induced by the random slopes. While our simulations and empirical application employ the same items at each time point, the model could also be applied to data in which only a subset of linking items were administered at each time point.

We generate 500 data sets for each parameterization with an initial single (i.e., unidimensional) normally distributed latent trait and equal and time-invariant item discriminations and fit two EIRMs to each, one with random intercepts for items (i.e., Equation 4 with $\sigma^2_{\zeta_1}$ fixed to 0) and another with random intercepts and slopes for items (i.e., Equation 4), resulting in 3,000 datasets and 6,000 models in total, and collect the model output for further analysis. We use the `glmer` function from the `lme4` R package to fit each EIRM as a generalized linear mixed model with a logit link function and cross-classified random effects for persons and items (Bates, et al., 2015; Gilbert, 2023). We use Wald tests to assess the statistical significance of the fixed effects and likelihood ratio tests to assess the significance of random effects or groups of fixed effects. We examine parameter bias and the calibration of the model standard errors. A detailed replication toolkit is available for researchers interested in extending the simulation or analysis of empirical data.

Bias

Figure 1 presents the bias for the time main effect (β_1), the treatment by time interaction effect (β_3), and the three-way interaction between treatment, time, and subscale (β_7). We see that for the two- and three-way interaction terms the item random intercepts specification (labelled RI in the figure) results in an increasing downward bias as IPD increases. When IPD is high (σ_{ζ_1} = 0.20 logits), we see that the downward bias is most severe, but still relatively small in magnitude. This downward bias is consistent with known properties of logistic regression that result in downwardly biased point estimates due to unobserved heterogeneity (e.g., omitted variables or a mis-specified model), even when the unobserved heterogeneity is independent of the variables in question, a property not shared by linear regression with continuous outcomes (Mood, 2010; Gilbert & Miratrix, 2023). The downward bias is not present when the true effect is precisely 0 because the downward bias is proportional to the true value (Mood, 2010, pp. 68-69). We emphasize that the observed bias is not due to shrinkage induced by empirical Bayes estimation of random effects that we would see in a two-step analysis that first fits a measurement model and then analyzes the resultant scores in a separate step (Soland, et al., 2022; Gilbert, 2024a, 2024b; Hedges, 1981). As a latent variable model where the measurement and regression models are simultaneously estimated, the EIRM does not suffer from such attenuation bias in general. Rather, the bias emerges from the misspecification of the model by omitting the relevant random effect term (Hox, et al., 2017).

[Insert Figure 1 Here]

Standard Error Calibration

Figure 2 displays the mean model-based SEs of the same three fixed effects as a percentage of the true SEs (i.e., the standard deviation of the point estimates). If the model-based SEs are well calibrated, we would expect them to fall on the horizontal line at 100%. While the SEs for the two- and three-way interaction terms are generally well calibrated across all models, falling within 10 percentage points of their true value, the SEs for the main effect of time become severely underestimated when IPD is high in the random intercepts model that constrains $\sigma^2_{\zeta_1}$ to 0. This occurs for the main effect of time only because when IPD is present, each finite draw of items will have a mean residual growth rate different from 0 due to sampling error, and when IPD is not modeled, the sample mean growth rate of the items is incorporated into the estimation of the average person growth rate β_1, creating greater variability across different samples of items. In other words, the SE for growth rate in the random intercepts model does not adjust for the additional uncertainty due to the selection of items onto the test. We can estimate the inflation of the SE due to IPD using the techniques of Generalizability Theory (Brennan, 1992) with the following formula:

$$\widehat{SE}(\beta_1)_{IPD} = \sqrt{\widehat{Var}(\beta_1)_{RI} + \frac{\sigma^2_{\zeta_1}}{I}} \quad (5)$$

where I is the number of items and $\widehat{Var}(\beta_1)_{RI}$ is the variance of the time slope from the random intercepts model. Clearly, when $\sigma^2_{\zeta_1}$ is high and I is low, the inflation of the SE can be substantial. We do not find the same pattern in the interaction effects because, so long as IPD affects treatment and control groups equally and all students answer the same items (as is the case in the simulation), this additional variability is subtracted out in the interaction effects.

[Insert Figure 2 Here]

Empirical Application

For our empirical application, we examine immediate and delayed treatment effects of the Model of Reading Engagement (MORE) randomized controlled trial (RCT) intervention. The MORE content literacy intervention is designed to improve first to third-grade grade students' domain and content background and vocabulary knowledge in science and social studies that are critical to reading comprehension. The MORE curriculum emphasizes thematic lessons that focus on a single topic over consecutive weeks in a semester and provides an intellectual structure for helping young children connect new content learning and vocabulary to a general schema (Anderson & Pearson, 1984; Kintsch, 2009; Perfetti, 2007). In a recent longitudinal investigation of MORE (Kim, et al., 2024), 30 elementary schools were randomly assigned to either a treatment or control condition. In the treatment condition, students participated in MORE content literacy lessons from Grades 1 to 3 during the school year and wide reading of thematically related informational texts in the summer following Grades 1 and 2. In the control condition, students participated in MORE lessons in only Grade 3. At the end of Grade 3, there were positive impacts on both researcher-designed domain specific reading comprehension tests in science (ES = 0.14) and state standardized end-of-grade domain general reading comprehension tests (ES = 0.11). An open question, however, is whether the full Grade 1 to 3 intervention fosters growth in vocabulary—a key malleable and potentially causal mechanism—compared to the partial Grade 3 intervention. This study provides an ideal context to address this question because students completed researcher-designed assessments in Grade 2 spring. Then, at the end of Grade 3, students completed another researcher-designed vocabulary test which included a repeated administration of the same vocabulary words tested at the end of Grade 2. Thus, we can estimate the immediate impact of MORE on the subset of students ($n = 1225$) who completed both second- and third-grade vocabulary tests and whether any treatment effect on vocabulary achievement persists, grows, or declines over a 12-month follow-up period[2].

The researcher-designed assessment of vocabulary knowledge depth includes 12 items. Each item lists a target word and prompts students to select the two words out of four choices that best go with the target word. For example, one item prompts students to "choose the two words that best go with the word **carnivore**" and the options were "fruit", "care", "meat", and "prey", of which the last two are the correct responses. Each item was scored dichotomously as correct (1) if students selected the two correct words, or incorrect (0) for any other response pattern. We apply the dichotomous scoring procedure rather than a partial credit or ordinal scoring system to match the approach of the original authors. The 12 vocabulary items included seven vocabulary words explicitly taught through the MORE intervention lessons ("taught words") and five conceptually related words that were not explicitly taught but were included in the lesson materials and activities such as read-alouds ("untaught words") and thus represented a farther degree of transfer from the

[2]The technically oriented reader might notice that typically, random slopes longitudinal models are not identified with only two time points because each subject's individual trajectory can be "perfectly" fit by the model (Muthen, 2000). The issue of non-identifiability does not apply here because the cross-classified structure of the data is additive, not multiplicative. That is, there is no interaction between the person and item random effects because such an interaction would be confounded with the error term, whereas the additive case allows for imperfect fit. Thus, such models may provide additional utility in empirical applications when only two time points are available. See O'Connell, et al. (2022, pp. 170-171), Hox, et al. (2017), and Shi, et al. (2010) for a discussion and additional references.

MORE curriculum (Barnett & Ceci, 2002). Here, we restrict our analysis to the subset of students (n = 1225) who completed the assessment in both Grades 2 and 3, and the subset of items that were included on both assessments (n = 12). The vocabulary assessment instrument and psychometric analyses at each time point are included in the Online Supplemental Materials (OSM), which show that the assessment had internal consistencies of 0.81 (G2) and 0.80 (G3), moderately to highly positive item discrimination parameters, and CFA revealed adequate fit of a unidimensional model at both pretest (CFI = 0.96, RMSEA = 0.04, SRMR = 0.030) and posttest (CFI = 0.98, RMSEA = 0.027, SRMR = 0.024)[3].

To explore immediate and delayed impacts of MORE on vocabulary knowledge depth, we fit four models, all including time, treatment, and their two-way interaction: (1) random intercepts for persons and items (analogous to Equation 2 with $\sigma^2_{\theta_1}$ constrained to 0) as a baseline, (2) random slopes for persons, random intercepts for items (analogous to Equation 2), (3) random slopes for persons and items (Equation 3), and (4) random slopes for persons and items with two- and three-way interaction effects (Equation 4). Because MORE was a cluster-randomized trial, we include school random effects in all models, but for clarity we omit them from the equation below. We only display the equation for Model 4 as all prior models are nested within it.

[3] We estimated the CFA model using the `lavaan` program in R (Rosseel, 2012) using the default estimation options, allowing for variable factor loadings by item. We treated the items as continuous because `lavaan` does not allow for logistic link functions and to obtain the standard fit statistics available in CFA models of continuous indicators. Furthermore, the assessment also included vocabulary words that were taught in Grade 1 MORE lessons and were tested in both Grade 1, halfway through the intervention, and in Grade 3, one year after the conclusion of the intervention. An analogous analysis of these words is included in the OSM and shows a similar pattern of results to the Grade 2 words analyzed here.

$$
\begin{aligned}
logit\left(P\left(y_{tij}=1\right)\right) \\
&= \beta_0 + \beta_1 time_{tij} + \beta_2 treat_j + \beta_3 treat \times time_{tij} \\
&+ \beta_4 baseline_{ij} + \beta_5 taught_i + \beta_6 treat \times taught_{ij} \\
&+ \beta_7 time \times taught_{tij} + \beta_8 treat \times time \times taught_{tij} + \theta_{0j} \\
&+ \zeta_{0i} + \theta_{1j} time_{tij} + \zeta_{1i} time_{tij}
\end{aligned}
\tag{6}
$$

$$
\begin{bmatrix} \theta_{0j} \\ \theta_{1j} \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma_{\theta_0}^2 & \rho_{10} \\ \rho_{01} & \sigma_{\theta_1}^2 \end{bmatrix})
$$

$$
\begin{bmatrix} \zeta_{0j} \\ \zeta_{1j} \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma_{\zeta_0}^2 & \tau_{10} \\ \tau_{01} & \sigma_{\zeta_1}^2 \end{bmatrix})
$$

Compared to Model 4, Model 3 omits $\beta_5, \beta_6, \beta_7$ and β_8, Model 2 omits $\sigma_{\zeta_1}^2$, and Model 1 omits $\sigma_{\theta_1}^2$.

In Model 4, the parameters of interest are $\sigma_{\zeta_1}^2$, quantifying IPD in the residual variance of item growth rates, β_2, the immediate treatment effect on untaught words in Grade 2, and β_3, the difference in the treatment effect on untaught words from Grade 2 to Grade 3 (i.e., potential fadeout for untaught words). $\beta_2 + \beta_3$ provides the delayed treatment effect on untaught words. We include baseline state test scores (standardized to mean 0 and unit variance), collected in Grade 1 winter as a covariate to improve the precision of the estimates (β_4). The baseline measure 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. We also include a main effect for taught words capturing differences in item easiness between taught and untaught words for control students in Grade 2 (β_5), an interaction between treatment and taught words capturing the difference in treatment effects for taught and untaught words in Grade 2 (β_6), an interaction between time and taught words capturing the difference in easiness between taught and untaught words between Grades 2 and 3 for control students (β_7), and a three-way interaction between treatment, time, and taught words capturing the difference in the two-way interaction for time and taught words for treatment students (β_8). Analogous to prior models, β_0 provides the control group mean in Grade 2 on untaught words, β_1 provides the average growth rate for control students on untaught words, and the random effects variances $\sigma^2_{\theta_0}$, $\sigma^2_{\theta_1}$, and $\sigma^2_{\zeta_0}$ provide the variability of student intercepts in Grade 2, the variability of student growth rates, and the variability in item easiness in Grade 2, respectively.

Results

The fitted models are presented in Table 1. Model 1 shows a positive but not significant treatment effect at immediate posttest at the end of Grade 2 (β_2 = 0.12 logits, p > 0.05), and that the magnitude of the average treatment effect *grows* over time through the end of Grade 3 (β_3 = 0.13, p < 0.05), showing the persistence of the MORE treatment effect in contrast to many studies that demonstrate fadeout of effects over time (see Bailey, et al., 2017; Wan, et al., 2021). The coefficients for time and baseline scores are strong and statistically significant, indicating that control student proficiency increased from Grade 2 to Grade 3 ($\beta_1 = 0.51$, p < 0.001) and that students with higher baseline scores had higher proficiency (β_4 = 0.98, p < 0.001). Model 2 adds the random slope for persons, and we observe that individual trajectories are highly heterogeneous ($\sigma^2_{\theta_1}$= 0.92), the treatment by time interaction term is no longer statistically significant, and the SE for the time coefficient has increased substantially. Model 3 adds the random slopes for items, representing IPD, and we see that there is substantial IPD ($\sigma^2_{\zeta_1}$= 0.25). A likelihood ratio test reveals that Model 3 is a significantly better fit to the data than Model 2, suggesting that the IPD in the dataset is significant (χ^2= 217.3, p < 0.001). The variation in item level growth trajectories is depicted in Figure 3, showing the model implied trajectories for each vocabulary word (i.e., $\beta_1 + \zeta_{1i}$), for the average student. Following the simulation results, we see that the SE for the main effect of time drastically increases, as the SE for Model 3 incorporates the additional uncertainty of which items were selected for test administration. To attempt to explain the moderate level of IPD, Model 4 adds two- and three-way interactions between treatment, time, and whether the item tested an explicitly taught vocabulary word. We observe that the treatment by taught word interaction is significant, indicating that at the immediate posttest, the treatment effect was smaller on taught words than untaught words (β_6 = -0.20, $p < 0.05$), a finding that matches prior separate analyses of Grade 2 vocabulary scores (Kim, et al., 2023, 2024). The main effect for treatment indicates that the treatment effect at immediate posttest is statistically significant for untaught words (β_2 = 0.23, $p < .05$). The three-way interaction between treatment, time, and taught word is non-significant. However, the variance of the IPD term remains unchanged, suggesting that the interaction effects have not captured substantial systematic variation in item growth, and the great majority of IPD remains unexplained, a function of the idiosyncratic characteristics of each item.

Figure 4 shows predicted probabilities of a correct response for the typical student and item for each treatment condition and “taught word” item status. Visually, we see that the immediate treatment effect on untaught words persists through the 12 month follow up. These results suggest that instead of diminishing over time, the MORE intervention was successfully able to lay a foundation for learning that persisted for untaught (far transfer) vocabulary words in the 12 months following treatment. Importantly, the larger treatment effects on the untaught vocabulary words suggest that the MORE intervention was not “teaching to the test” and thus the results are unlikely to be attributable to score inflation (Koretz, 2005). Furthermore, in contrast to a two-step approach in which the four outcomes were modeled separately (e.g., G2 taught, G2 untaught, G3 taught, and

G3 untaught), the EIRM allows direct tests of *differences* in treatment effect size across these subscales in the parsimony of a single model (Gilbert, 2023).

[Insert Table 1, Figure 3, and Figure 4 Here]

Discussion

Analysis of individual growth in education has typically emphasized between-person predictors of growth through person characteristic by time interactions. When item-level data are available, another perspective is possible, namely, item characteristic by time interactions to assess the extent to which proficiency on different items or subscales may develop at different rates. In the educational measurement literature, changing item properties over time has been viewed as a nuisance under the rubric of IPD. In this study, we argue that IPD can represent substantively meaningful differential learning on different items or subscales, and the EIRM with a random slope for time at the item level provides an opportunity to better understand the facets of student growth if student learning is not constant across all items over time.

Results of the data simulation revealed that when a high degree of IPD is present in the data but ignored in the model, point estimates for interaction terms are slightly biased downward, and standard errors can be underestimated for main effects, but not interaction effects involving time. Therefore, researchers employing the EIRM should consider the possibility of IPD and test for its presence with a random slope for time at the item level, even if IPD is not of primary interest, to obtain accurate parameter estimates and SEs, particularly when examining interaction effects such as the time by treatment by subscale effects examined here. For example, as shown in the empirical application, when IPD is included in Model 3, the SE for the main effect of time increases dramatically, in line with the simulation results and Equation **Error! Reference source not found.**. The empirical application further showed that the MORE literacy intervention had persistent effects on student vocabulary ability on the far transfer untaught words from the end of treatment in Grade 2 through a 12-month follow up in Grade 3. While explicitly taught words were easier on average than untaught words, the treatment effect was larger on the more difficult untaught words, providing evidence that treatment students were successfully able to transfer their learning to new contexts.

To extend the applicability of the EIRM to more diverse applied contexts, the simple example of a unidimensional Rasch model employed in this study could be easily augmented to include varying item discriminations (i.e., a 2PL model; Rockwood & Jeon, 2019; Burkner, 2021; Gilbert, 2023), missing data (de Boeck et al., 2016), multidimensionality (de Boeck & Wilson, 2014), and non-dichotomous responses (Bulut, et al., 2021; Gilbert, Hieronymus, et al., 2024). While widely applicable, a potential limitation of the EIRM is the interpretation and communication of the results. Log-odds may be more difficult to explain and justify to practitioners than a more familiar sum or scaled score. Previous studies of the EIRM suggested two approaches to increase the communicability of the results (Gilbert, Kim, & Miratrix, 2023). First, fitted models can be used to estimate population average probabilities at each time point, as depicted in Figure 4, for example by using the R package `ggeffects` (Lüdecke, 2018). Second, treatment effects on the logit scale can be converted into a Cohen's *d* type effect size by "y-standardization" (see Breen, Karlson, & Holm, 2018 for the single-level case; see Hox, Moerbeek, & Van de Schoot, 2017, Chapter 6 for the multilevel case), whereby the logit-scale coefficient β_{logit} is divided by the estimated total standard deviation of a latent continuous variable Y* that gives rise to the observed dichotomous response Y, using the following formula

$$\beta_{ystd} = \frac{\beta_{logit}}{SD(Y^*)} = \frac{\beta_{logit}}{\sqrt{\frac{\pi^2}{3} + \sigma^2_{\theta_0} + \sigma^2_F}}$$

in which $\frac{\pi^2}{3} = 3.29$ is the variance of the logistic distribution, $\sigma^2_{\theta_0}$ is the variance of the person intercepts at baseline, and σ^2_F is the variance of the fixed effects (i.e., the variance of the estimated linear predictor on the logit scale). The y-standardized coefficients could then be compared to other metrics or used in meta-analysis. In the context of this study, for example, the estimated standardized effect size on untaught vocabulary words at immediate posttest is equal to

$$\beta_{ystd} = \frac{\beta_{logit}}{SD(Y^*)} = \frac{.23}{\sqrt{3.29 + .93 + 2.6}} = 0.09$$

a small but significant positive impact. Such an effect size could then be converted to a percentile gain (about 3.3 percentile points, see Hippel, 2023), or an approximate number of additional items answered correctly.

Another challenge of the longitudinal EIRM approach proposed here is that the many parameters of the model require large samples of both items and persons for consistent estimation. For example, simulation studies by Soland, et al. (2022) show that IRT-based scoring methods can yield substantial bias in estimated treatment effects with short, 4-item scales, though the biases are substantially reduced as the number of items increases to 12 or more. Similarly, simulations in Gilbert, Kim, and Miratrix (2023) show that achieving 80% statistical power to detect random slope treatment effect variation at the item level is reached at 300 subjects, 20 items, and relatively large random slope standard deviations of .40. Additional studies have confirmed that EIRM approaches successfully recover item parameters and regression coefficients when sample sizes are large (Gilbert, Kim, & Miratrix, 2023, p. 896). As such, our approach is best suited for relatively large-scale data analysis contexts.

In conclusion, item parameter drift has traditionally been considered a nuisance in the educational measurement literature, but it has the potential to provide substantive insight into the learning process as proficiency on different items may develop at different rates. By explicitly modeling IPD, the EIRM allows for more nuanced and fine-grained insights into the nature of student learning over time. In particular, the IPD model may provide a more generalizable perspective on student growth by incorporating the uncertainty of item selection into the standard errors of the growth estimates. Such generalizability is particularly important in domains such as vocabulary, in which the underlying construct can never be fully measured by any finite set of items. Researchers can use such insights to provide more actionable information to stakeholders and better understand the ways in which individual growth is a multifaceted phenomenon.

References

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 255-291). Routledge.

Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., Rozelle, S., & Domingue, B. (2023). *Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials* (EdWorkingPaper No. 23-754). Annenberg Institute at Brown University. https://doi.org/10.26300/1nw4-na96

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7-39.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Binici, S. (2007). *Random-effect differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: a comparison of estimation methods*. [Doctoral dissertation, The Florida State University]. DigiNole: FSU's Digital Repository. https://diginole.lib.fsu.edu/islandora/object/fsu:182005

Breen, R., Karlson, K. B., & Holm, A. (2018). Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, *44*, 39-54.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27-34.

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

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

Cho, S. J., Athay, M., & Preacher, K. J. (2013). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 353-381.

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

De Boeck, P., & Wilson, M. (2014). Multidimensional explanatory item response modeling. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: applications to typical performance assessment* (pp. 252-271). New York: Routledge.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1-28.

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

de Bock, E., Hardouin, J. B., Blanchin, M., Le Neel, T., Kubis, G., Bonnaud-Antignac, A., ... & Sebille, V. (2016). Rasch-family models are more valuable than score-based approaches for analysing longitudinal patient-reported outcomes with missing data. *Statistical Methods in Medical Research*, *25*(5), 2067-2087.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*(1), 33-51.

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*(4), 432-457.

Gamerman, D., Gonçalves, F. B., & Soares, T. M. (2017). Differential item functioning. In W. van der Linden (Ed.), *Handbook of item response theory. Volume three: Applications* (pp. 67-86). CRC Press.

Gilbert, J. B. (2023). Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in R. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-023-02245-8

Gilbert, J. B. (2024a). Estimating treatment effects with the explanatory item response model. *Journal of Research on Educational Effectiveness*. Advance online publication. https://doi.org/10.1080/19345747.2023.2287601

Gilbert, J. B. (2024b). *How measurement affects causal inference: Attenuation bias is (usually) more important than scoring weights* (EdWorkingPaper No. 23-766). Annenberg Institute at Brown University. https://doi.org/10.26300/4hah-6s55

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, 48*(6), 889-913.

Gilbert, J. B., & Miratrix, L. W. (2023, March 6). Recovering effect sizes from dichotomous variables using logistic regression. *CARES Lab Blog*. https://cares-blog.gse.harvard.edu/post/logistic-effects/

Gilbert, J. B., Hieronymus, F., Eriksson, E., & Domingue, B. W. (2024). *Item-level heterogeneous treatment effects of selective serotonin reuptake inhibitors (SSRIs) on depression: Implications for inference, generalizability, and identification*. arXiv. https://arxiv.org/abs/2402.04487

Gilbert, J. B., Miratrix, L. W., Joshi, M., & Domingue, B. (2024). Disentangling person-dependent and item-dependent causal effects: Applications of item response theory to the estimation of treatment effect heterogeneity. *Journal of Educational and Behavioral Statistics*. Advance online publication. https://doi.org/10.3102/10769986241240085

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107-128.

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

Jeon, M., & Rabe-Hesketh, S. (2016). An autoregressive growth model for longitudinal item analysis. *Psychometrika*, *81*(3), 830-850.

Kintsch, W. (2009). Learning and constructivism. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction: Success or failure?* (pp. 235-253). Routledge.

Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & 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

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology*, *113*(1), 3-26.

Kim, J. S., Gilbert, J. B., Relyea, J. E., Rich, R., Scherer, E., Burkhauser, M. A., & Tvedt, J. N. (2024). Time to transfer: long-term effects of a sustained and spiraled content literacy intervention in the elementary grades. *Developmental Psychology*. Advance online publication. https://doi.org/10.1037/dev0001710

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

Lee, H., & Geisinger, K. F. (2019). Item parameter drift in context questionnaires from international large-scale assessments. *International Journal of Testing*, *19*(1), 23-51.

Lee, W., & Cho, S. J. (2017). The consequences of ignoring item parameter drift in longitudinal item response models. *Applied Measurement in Education*, *30*(2), 129-146.

Liu, S., Kuppens, P., & Bringmann, L. (2021). On the use of empirical Bayes estimates as measures of individual traits. *Assessment*, *28*(3), 845-857.

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486.

Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772.

Luo, J., Wang, M. C., Ge, Y., Chen, W., & Xu, S. (2020). Longitudinal invariance analysis of the short grit scale in Chinese young adults. *Frontiers in Psychology*, *11*, 1-9.

Luo, S., Zou, H., Stebbins, G. T., Schwarzschild, M. A., Macklin, E. A., Chan, J., ... & members of Parkinson Study Group SURE-PD3 Investigators. (2022). Dissecting the domains of Parkinson's disease: insights from longitudinal item response theory modeling. *Movement Disorders*, *37*(9), 1904-1914.

te Marvelde, J. M., Glas, C. A., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, *66*(1), 5-34.

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

Montoya, A. K., & Jeon, M. (2020). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement*, *44*(2), 118-136.

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

Muthén, B. (2000). Methodological issues in random coefficient growth modeling using a latent variable framework: applications to the development of heavy drinking ages 18–37. In J. S. Rose, L. Chassin, C. C. Presson, & S. J. Sherman (Eds.), *Multivariate applications in substance use research* (pp. 113-140). Psychology Press.

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.

O'Connell, A. A., McCoach, D. B., & Bell, B. A. (Eds.). (2022). *Multilevel modeling methods with introductory and advanced applications*. Information Age Publishing.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement*, *30*(2), 100-120.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357-383.

Proust-Lima, C., Philipps, V., Perrot, B., Blanchin, M., & Sébille, V. (2022). Modeling repeated self-reported outcome data: A continuous-time longitudinal item response theory model. *Methods*, *204*, 386-395.

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Randall, J., Cheong, Y. F., & Engelhard Jr, G. (2011). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement*, *71*(1), 129-147.

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

Rockwood, N. J., & Jeon, M. (2019). Estimating complex measurement and growth models using the R package PLmixed. *Multivariate Behavioral Research*, *54*(2), 288-306.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63-84.

Sales, A., Prihar, E., Heffernan, N., & Pane, J. F. (2021). The effect of an intelligent tutor on performance on specific posttest problems [Paper presentation]. *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, Online. https://eric.ed.gov/?id=ED615618

Shi, Y., Leite, W., & Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. *British Journal of Mathematical and Statistical Psychology*, *63*(1), 1-15.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Soland, J., Kuhfeld, M., & Edwards, K. (2022). How survey scoring decisions can influence your study's results: A trip through the IRT looking glass. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000506

Sukin, T. M. (2010). *Item parameter drift as an indication of differential opportunity to learn: An exploration of item flagging methods & accurate classification of examinees*. [Doctoral dissertation, University of Massachusetts Amherst]. Scholarworks @ UMass Amherst. https://scholarworks.umass.edu/open_access_dissertations/301/

Ten Have, T. R., & Localio, A. R. (1999). Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics*, *55*(4), 1022-1029.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443-464.

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

von Hippel, P. T. (2023). *Multiply by 37: A surprisingly accurate rule of thumb for converting effect sizes from standard deviations to percentile points* (EdWorkingPaper No. 23-829). Annenberg Institute at Brown University. https://doi.org/10.26300/xk0b-ft25

Waclawiw, M. A., & Liang, K. Y. (1994). Empirical Bayes estimation and inference for the random effects model with binary response. *Statistics in Medicine*, *13*(5-7), 541-551.

Wang, C., & Nydick, S. W. (2020). On longitudinal item response theory models: A didactic. *Journal of Educational and Behavioral Statistics*, *45*(3), 339-368.

Wan, S., Bond, T. N., Lang, K., Clements, D. H., Sarama, J., & Bailey, D. H. (2021). Is intervention fadeout a scaling artefact? *Economics of Education Review*, *82*, 102090.

Wilson, M., Zheng, X., & McGuire, L. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement*, *13*(1), 1-22.

Ye, F. (2016). Latent growth curve analysis with dichotomous items: Comparing four approaches. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 43-61.

Figure 1. Parameter Bias by Method

(RI = Random Intercepts for Items, RS = Random Slopes for Items)

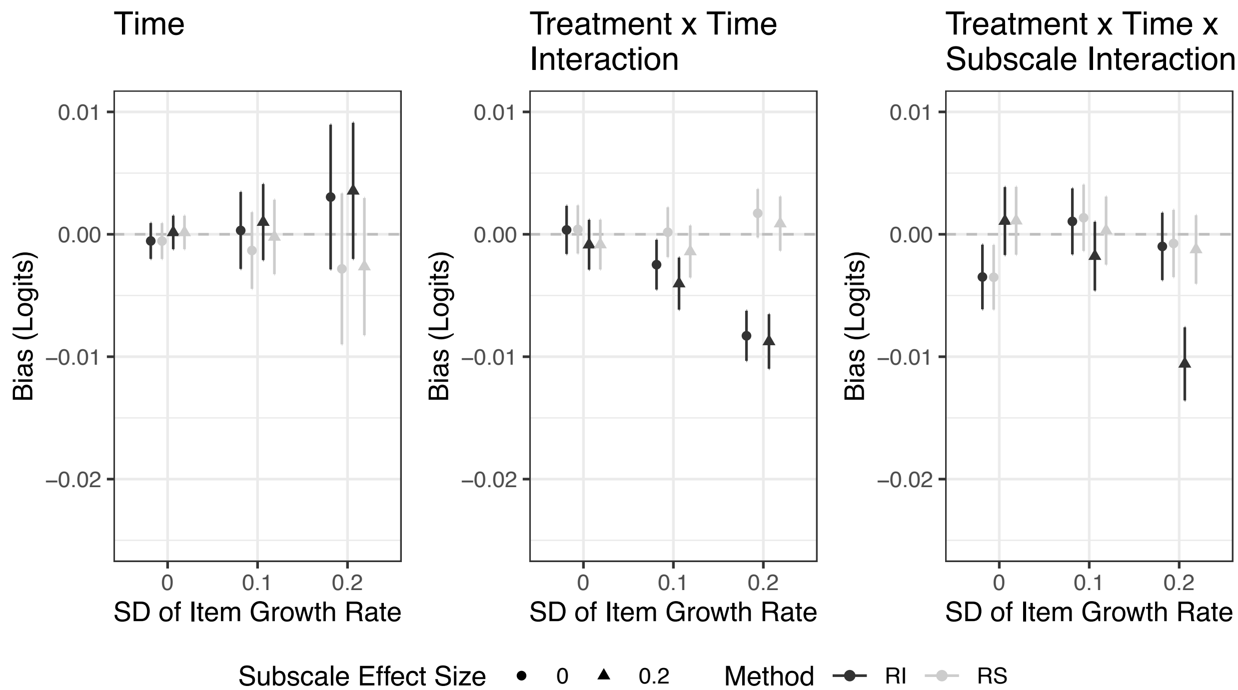


Figure 2. Standard Error Calibration by Method

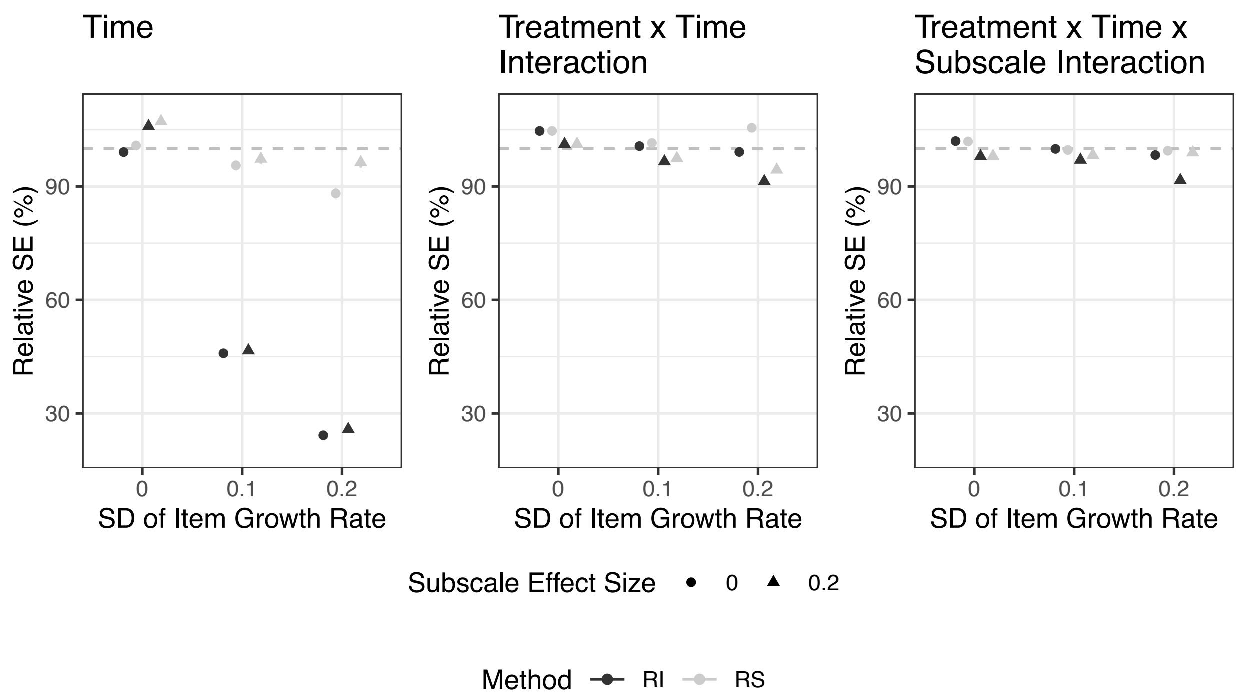
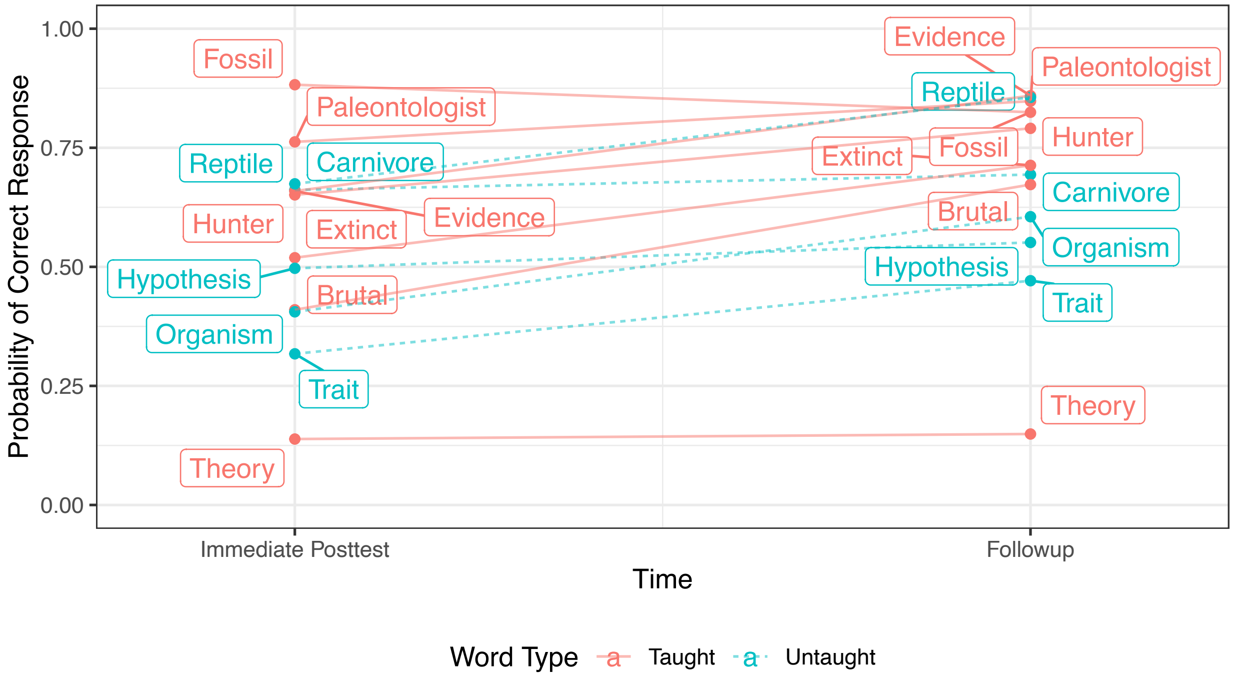


Figure 3. Item-Level Growth Trajectories Derived from Model 3



Notes. A table of item-specific growth rates and a graph of empirical Bayes estimates for each item's residual growth rate and a 95% confidence interval is presented in our online supplemental materials.

Figure 4. Prototypical Probabilities of a Correct Response for Treatment and Control Students on Taught and Untaught Vocabulary Words Derived from Model 4

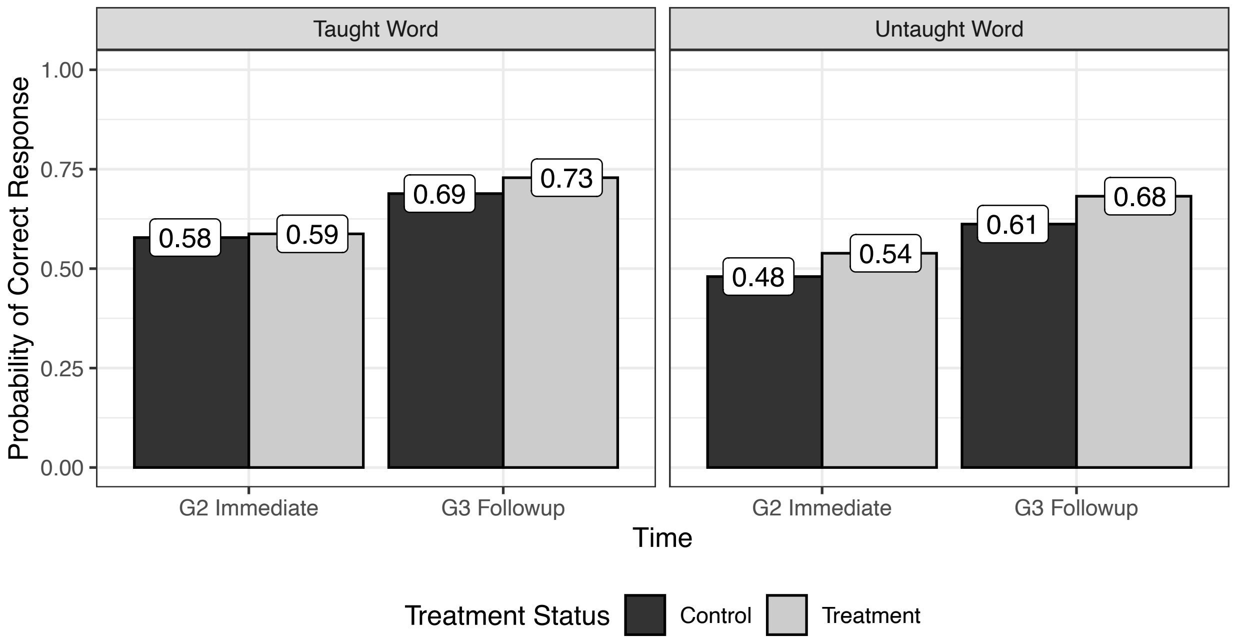


Table 1. EIRM Results for the MORE Intervention Data

	M1		M2		M3		M4	
Predictors	*Log-Odds*	*SE*	*Log-Odds*	*SE*	*Log-Odds*	*SE*	*Log-Odds*	*SE*
Constant (β_0)	0.04	0.28	0.03	0.29	0.05	0.30	-0.18	0.45
Treatment (β_2)	0.12	0.09	0.12	0.10	0.12	0.10	0.23 *	0.11
Time (β_1)	0.51 ***	0.04	0.52 ***	0.06	0.50 **	0.16	0.54 *	0.24
Baseline MAP (β_4)	0.98 ***	0.03	1.00 ***	0.03	1.01 ***	0.03	1.01 ***	0.03
Treatment X Time (β_3)	0.13 *	0.06	0.12	0.08	0.12	0.08	0.07	0.10
Taught Word (β_5)							0.39	0.59
Treatment X Taught (β_6)							-0.20 *	0.08
Time X Taught (β_7)							-0.06	0.31
Treatment X Time X Taught (β_8)							0.08	0.12
Random Effects								
Scale Variance	3.29		3.29		3.29		3.29	
$\sigma^2_{\theta_0}$(Student)	0.56		0.90		0.93		0.93	
σ^2_{v} (School)	0.04		0.04		0.04		0.04	
$\sigma^2_{\zeta_0}$ (Item)	0.89		0.97		1.01		0.99	
$\sigma^2_{\theta_1}$ (Student Growth)			0.92		0.96		0.96	
$\sigma^2_{\zeta_1}$ (IPD)					0.25		0.25	
ρ_{01} (Student Corr.)			-0.57		-0.57		-0.57	
τ_{01} (Item Corr.)					-0.19		-0.19	
N Students	1225		1225		1225		1225	
N Items	12		12		12		12	
N Schools	29		29		29		29	
Observations	29327		29327		29327		29327	
Deviance	30566.468		30264.879		30047.539		30039.239	

* *p<0.05* ** *p<0.01* *** *p<0.001*

Notes. MAP = Measure of Academic Progress, our measure of baseline ability.

Acknowledgements

The authors wish to thank Douglas Mosher, Jackie Relyea, and the three anonymous reviewers for their helpful comments on this manuscript.

Data and Code Availability: A detailed replication toolkit containing the data and code used in this study (as well as some supplemental analyses) is available at the following URL: https://doi.org/10.7910/DVN/ZF1LKZ

Appendix: Sample R Code to Fit the Longitudinal EIRM

The code below shows the basic R syntax to fit various longitudinal EIRMs with dichotomous outcome `correct`, treatment indicator `treat`, time variable `time`, item type indicator `itemtype`, student identifier `s_id`, and item identifier `item`. For clarity, we omit `family = binomial` and `data = dataset` from each `glmer` function call.

```
# load the relevant library
library(lme4)

# baseline model:
# random intercepts with treatment by time interaction
glmer(correct ~ treat*time + (1|s_id) + (1|item))

# add random slopes for persons
glmer(correct ~ treat*time + (time|s_id) + (1|item))

# add random slopes for items (IPD)
glmer(correct ~ treat*time + (time|s_id) + (time|item))

# add interaction between treatment, time, and itemtype
glmer(correct ~ treat*time*itemtype + (time|s_id) + (time|item))
```